METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

A Comparative Study of Approximations for Perturbation Analysis of Principal Components

Jacques Bénasséni [1], Alain Mom [1]

[1] *Université Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France.*

Methodology, 2025, Vol. 21(1), 27–45, https://doi.org/10.5964/meth.15357

Received: 2024-08-20 • **Accepted:** 2025-02-07 • **Published (VoR):** 2025-03-31

Handling Editor: Tamás Rudas, Eötvös Loránd University, Budapest, Hungary

Corresponding Author: Jacques Bénasséni, Université Rennes 2, Place du Recteur Henri Le Moal CS 24307 - 35043 Rennes cedex, France. E-mail: jacques.benasseni@univ-rennes2.fr

Abstract

Principal component analysis is a well known method for dimension reduction based on the covariance matrix associated to a multivariate data table. Therefore, a large amount of work has been devoted to analyzing the sensitivity of the eigenstructure of this matrix to influential observations. In order to evaluate the effect of deleting one or a small subset of observations, several approximations for the perturbed eigenelements have been proposed. This paper provides a theoretical and numerical comparison of the main approximations. A special emphasis is given to those based on Rayleigh quotients since they are under-utilized given their excellent performance. A general approach, using refined inequalities, is proposed in order to get a precise evaluation of their accuracy without having to recompute the exact perturbed eigenvalues and eigenvectors. This approach is of specific interest from a computational standpoint. Theoretical developments are illustrated with a numerical study which emphasizes the accuracy of approximations based on Rayleigh quotients.

Keywords

approximation, eigenvalue and eigenvector, covariance matrix, principal component analysis, perturbation, Rayleigh quotient.

We consider a $n \times p$ data matrix $\mathbf{X}$ in which the n rows are the observation vectors $\mathbf{x}_i^t \in \mathbb{R}^p$, $i = 1, \ldots, n$. Letting $\overline{\mathbf{x}} = (1/n)\sum_{i=1}^{n} \mathbf{x}_i$ denote the mean vector of dimension p, the covariance matrix $\mathbf{S} = (1/n)\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t$ is involved in a wide range of statistical methods including principal component analysis or multiple regression for example. Since $\mathbf{S}$ is known to be highly prone to influential observations, sensitivity aspects for principal component analysis have been discussed in several papers including Critchley (1985), Jolliffe (2002), Pack et al. (1988), Prendergast (2008), Prendergast and Li Wai Suen (2011), Tanaka (1988) among many others. It should be noted that perturbation issues for this method are still the subject of active research as emphasized by the recent work of Masioti et al. (2023) for example. In order to assess the influence of a small subset I of r observations on the eigenstructure of $\mathbf{S}$, a possible approach consists of studying the effect of removing these r observations on the eigenelements of $\mathbf{S}$. If we note a significant modification of the eigenelements, we know that these observations are strongly influential on the results. In this framework, letting $\widetilde{\mathrm{S}}$ denote the covariance matrix obtained without the subset of observations indexed by I, several authors have studied the relationship between the eigenelements of $\mathbf{S}$ and those of $\widetilde{\mathrm{S}}$. More specifically, providing approximations to the eigenelements of $\widetilde{\mathrm{S}}$ allows to detect influential subsets of observations without having to recompute the exact modified eigenvalues and eigenvectors. Hadi and Nyquist (1993), Masioti et al. (2023), and Wang and Nyquist (1991) study the effect of deleting a single observation while Bénasséni (2018), Enguix-González et al. (2005) and Wang and Liski (1993) focus on the general case where I comprises several observations. In these works, approximations to the eigenvalues and eigenvectors are obtained by retaining the first terms in power expansions of these parameters. Independently of these works, Bénasséni (1987) suggests using approximations based on Rayleigh quotients together with inequalities given in Wilkinson (1988). More generally, approximations for the covariance matrix are also the subject of research in a wider framework as emphasized by other works including for example Enguix-González et al. (2015) which considers the moments of the eigenelements or Enguix-González et al. (2012) dealing with the conditional bias of eigenvalues.

The contribution of this work is twofold. First, some new elements of comparison are provided between approximations derived from power expansions and those based on Rayleigh quotients. This comparison is based on some simple theoretical relations together with the numerical study of a data table which is intended to provide some guidance on the approximations to choose in practice. Second, refined inequalities in Chatelin (2012) are introduced in order to evaluate the accuracy of approximations without having to recompute the eigenelements of $\widetilde{\mathrm{S}}$. It is proved that these inequalities always provide a sharper evaluation than those used in Bénasséni (1987). This is a definite advantage from a computational standpoint.

In this work, the eigenvalues $\lambda_1 > \lambda_2 \ldots > \lambda_p \geq 0$ of $\mathbf{S}$ are assumed simple and associated to the normalized eigenvectors $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots \boldsymbol{\phi}_p$. In the same way, the eigenvalues $\widetilde{\lambda}_1 > \widetilde{\lambda}_2 \ldots > \widetilde{\lambda}_p \geq 0$ of the perturbed matrix $\widetilde{\mathrm{S}}$ are also assumed simple and associated to the eigenvectors $\widetilde{\boldsymbol{\phi}}_1, \widetilde{\boldsymbol{\phi}}_2, \ldots \widetilde{\boldsymbol{\phi}}_p$.

Approximations Based On Power Expansions

Theoretical Background On Matrix Perturbations

Referring to Bénasséni (2018) or Enguix-González et al. (2005) and letting $\overline{\mathbf{x}}_I = (1/r)\sum_{i \in I}\mathbf{x}_i$, we know that, when a subset of observations indexed by I is deleted, the covariance matrix $\mathbf{S}$ is transformed to $\widetilde{\mathrm{S}}$ which can be expressed as:

$$\widetilde{\mathrm{S}} = \mathrm{S} + \left(\frac{r}{n-r}\right)\left[\mathrm{S} - \frac{1}{r}\sum_{i \in I}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t\right] + \left(\frac{r}{n-r}\right)^2\left[-(\overline{\mathbf{x}}_I - \overline{\mathbf{x}})(\overline{\mathbf{x}}_I - \overline{\mathbf{x}})^t\right] \quad (1)$$

We then have a perturbation of the form $\widetilde{\mathrm{S}} = \mathrm{S} + \epsilon\mathbf{M} + \epsilon^2\mathbf{N}$ with $\epsilon = \frac{r}{n-r}$, $\mathbf{M} = \mathrm{S} - (1/r)\sum_{i \in I}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t$ and $\mathbf{N} = -(\overline{\mathbf{x}}_I - \overline{\mathbf{x}})(\overline{\mathbf{x}}_I - \overline{\mathbf{x}})^t$. Following matrix perturbation theory detailed in Wilkinson (1988) for example or referring to Sibson (1979), we know that, if ϵ is sufficiently small, for each simple eigenvalue λ of $\mathbf{S}$ there is an eigenvalue $\widetilde{\lambda}$ of $\widetilde{\mathrm{S}}$ given by a convergent power series:

$$\widetilde{\lambda} = \lambda + \gamma_1\epsilon + \gamma_2\epsilon^2 + \ldots + \gamma_m\epsilon^m + O(\epsilon^{m+1}) \quad (2)$$

with a corresponding eigenvector which can also be expressed under a convergent power series:

$$\widetilde{\boldsymbol{\phi}} = \boldsymbol{\phi} + \boldsymbol{\psi}_1\epsilon + \boldsymbol{\psi}_2\epsilon^2 + \ldots + \boldsymbol{\psi}_m\epsilon^m + O(\epsilon^{m+1}) \quad (3)$$

The parameters $\gamma_1, \gamma_2, \ldots, \gamma_m$ and $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_m$ are derived by equating the coefficients of $\epsilon, \epsilon^2, \ldots, \epsilon^m$ in the equation $\widetilde{\mathrm{S}}\widetilde{\boldsymbol{\phi}} = \widetilde{\lambda}\widetilde{\boldsymbol{\phi}}$. It should also be noted that the perturbation $\widetilde{\lambda}_k$ of λ_k may not necessarily be the kth largest eigenvalue of $\widetilde{\mathrm{S}}$ if the subset of observations indexed by I has initially a strong influence on λ_k for example. In this case, following Critchley (1985), one can simply assume that the eigenvalues $\widetilde{\lambda}_k$ have been reordered in decreasing order and that their corresponding eigenvectors $\widetilde{\boldsymbol{\phi}}_k$ have been relabeled. However, the reader is referred to Masioti et al. (2023) for a more comprehensive discussion on this topic.

Formulae for the Approximations Derived From Power Expansions

For any integer $m \geq 1$, retaining only terms of order lower or equal to m in ϵ in (2) and (3) provides the approximations $\widetilde{\lambda}^{(m)}$ and $\widetilde{\boldsymbol{\phi}}^{(m)}$ of order m for the eigenvalues and eigenvectors of $\widetilde{\mathrm{S}}$.

This is the general approach suggested in Bénasséni (2018), Enguix-González et al. (2005) and Wang and Liski (1993). Assuming that $\epsilon = \frac{r}{n-r}$ is sufficiently small to ensure the convergence of the above power series, Enguix-González et al. (2005) provide the following approximations for $k = 1, \ldots, p$:

$$\widetilde{\lambda}_k^{(1)} = \lambda_k + \left(\frac{r}{n-r}\right)\left(\lambda_k - \frac{1}{r}\sum_{i \in I} \alpha_{ki}^2\right) \tag{4}$$

$$\widetilde{\lambda}_k^{(2)} = \widetilde{\lambda}_k^{(1)} + \left(\frac{r}{n-r}\right)^2 \left[-\alpha_{kI}^2 - \sum_{j \neq k} \frac{1}{\lambda_j - \lambda_k}\left(\frac{1}{r}\sum_{i \in I} \alpha_{ki}\alpha_{ji}\right)^2\right] \tag{5}$$

$$\widetilde{\boldsymbol{\phi}}_k^{(1)} = \boldsymbol{\phi}_k + \left(\frac{r}{n-r}\right)\sum_{j \neq k}\left(\frac{1}{r}\sum_{i \in I} \alpha_{ki}\alpha_{ji}\right)\frac{\boldsymbol{\phi}_j}{\lambda_j - \lambda_k} \tag{6}$$

where $\alpha_{ki} = \phi_k^t(\mathbf{x}_i - \overline{\mathbf{x}})$, for $i = 1, \ldots, n$ and $\alpha_{kI} = (\sum_{i \in I} \alpha_{ki})/r$.

Details on the derivation of the above expressions are given in Bénasséni (2018). The reader is referred to Enguix-González et al. (2005) for the formula of $\widetilde{\boldsymbol{\phi}}_k^{(2)}$ which is fairly long and therefore omitted in this paper. It should also be pointed out that a comprehensive study of approximations for the unbiased matrix $(n-r)\widetilde{\mathrm{S}}/(n-r-1)$ is provided in this last reference. In the remaining of the paper, we define also the approximations of order zero as $\widetilde{\lambda}^{(0)} = \lambda$ and $\widetilde{\boldsymbol{\phi}}^{(0)} = \boldsymbol{\phi}$ for notational convenience.

Finally, when studying the influence of a single observation, note that approximations are simply obtained by taking $r = 1$, $I = \{i\}$ and $\epsilon = \frac{1}{n-1}$ in the previous developments. The reader more specifically interested by this case will refer to a series of papers including Critchley (1985), Hadi and Nyquist (1993), Masioti et al. (2023) and Wang and Nyquist (1991).

Approximations Based On Rayleigh Quotients

Rayleigh Quotients as Approximations to the Perturbed Eigenvalues

Assuming that weights are given to the observations, Bénasséni (1987) studies the effects of modifying these weights on the eigenvalues and eigenvectors of the covariance matrix. Deleting a small subset of observations indexed by I is therefore a particular case of his approach which consists simply of modifying to zero the corresponding weights. As approximations to $\widetilde{\lambda}_k$ for $k = 1, \ldots, p$, this author suggests using the Rayleigh quotient $q_k^{(0)} = \left(\widetilde{\boldsymbol{\phi}}_k^{(0)}\right)^t \widetilde{\mathrm{S}}\widetilde{\boldsymbol{\phi}}_k^{(0)} / \left(\widetilde{\boldsymbol{\phi}}_k^{(0)}\right)^t \widetilde{\boldsymbol{\phi}}_k^{(0)}$ for $\widetilde{\mathrm{S}}$ and the initial normalized eigenvector $\widetilde{\boldsymbol{\phi}}_k^{(0)}$, and the Rayleigh quotient $q_k^{(1)} = \left(\widetilde{\boldsymbol{\phi}}_k^{(1)}\right)^t \widetilde{\mathrm{S}}\widetilde{\boldsymbol{\phi}}_k^{(1)} / \left(\widetilde{\boldsymbol{\phi}}_k^{(1)}\right)^t \widetilde{\boldsymbol{\phi}}_k^{(1)}$ for $\widetilde{\mathrm{S}}$ and the approximation of order one $\widetilde{\boldsymbol{\phi}}_k^{(1)}$ to $\widetilde{\boldsymbol{\phi}}_k$.

Error Analysis

From a computational standpoint, it is of major interest to evaluate the accuracy of approximations without having to recompute the exact eigenelements of $\widetilde{\mathrm{S}}$. In order to do this, Bénasséni (1987) suggests using inequalities provided in Wilkinson (1988).

Focusing on $\widetilde{\lambda}_k$ and its corresponding eigenvector $\widetilde{\boldsymbol{\phi}}_k$ and, from now on, assuming without change of notation that $\widetilde{\boldsymbol{\phi}}_k^{(m)}$ has been normalized, let $\eta_k^{(m)} = \left\| \widetilde{S}\widetilde{\boldsymbol{\phi}}_k^{(m)} - q_k^{(m)}\widetilde{\boldsymbol{\phi}}_k^{(m)} \right\|_2$ for $m = 0, 1$ where $\| \,.\, \|_2$ stands for the two norm. Assume that $c_k^{(m)} \in \mathbb{R}^{+*}$ is a nonzero positive constant such that $\left|\widetilde{\lambda}_j - q_k^{(m)}\right| > c_k^{(m)}$ for $j = 1, \ldots, p$ with $j \neq k$. Then the accuracy of $q_k^{(m)}$ as approximation to $\widetilde{\lambda}_k$, and of $\widetilde{\boldsymbol{\phi}}_k^{(m)}$ as approximation to $\widetilde{\boldsymbol{\phi}}_k$, is analyzed in Bénasséni (1987) using the following inequalities given in (Wilkinson, 1988, pp. 172–176):

$$\left\|\widetilde{\boldsymbol{\phi}}_k - \widetilde{\boldsymbol{\phi}}_k^{(m)}\right\|_2^2 \leq \left(\frac{\eta_k^{(m)}}{c_k^{(m)}}\right)^2 \left[1 + \left(\frac{\eta_k^{(m)}}{c_k^{(m)}}\right)^2\right] \tag{7}$$

and if $\eta_k^{(m)} / c_k^{(m)} < 1$:

$$\left|\widetilde{\lambda}_k - q_k^{(m)}\right| \leq \left[\frac{\left(\eta_k^{(m)}\right)^2}{c_k^{(m)}}\right] \Big/ \left[1 - \left(\frac{\eta_k^{(m)}}{c_k^{(m)}}\right)^2\right] \tag{8}$$

Using $\left\|\widetilde{\boldsymbol{\phi}}_k\right\|_2 = \left\|\widetilde{\boldsymbol{\phi}}_k^{(m)}\right\|_2 = 1$, note that (7), can be written with the cosine between $\widetilde{\boldsymbol{\phi}}_k$ and $\widetilde{\boldsymbol{\phi}}_k^{(m)}$ as:

$$1 - \left[\left(\eta_k^{(m)}\right)^2 / 2\left(c_k^{(m)}\right)^2\right]\left[1 + \left(\eta_k^{(m)} / c_k^{(m)}\right)^2\right] \leq \cos\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(m)}\right). \tag{9}$$

In practice, it is necessary to give the parameter $c_k^{(m)}$ a value in the previous inequalities. When studying the influence of a single observation, it should be noted that (1) can be expressed as

$$\widetilde{S} = \frac{n}{n-1}\left[S - \frac{1}{n-1}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^t\right] \tag{10}$$

so that we have a rank one perturbation. In this case, the parameter $c_k^{(m)}$ is given a value in Bénasséni (1987) using bounds, derived from the Courant-Fischer theorem, for the eigenvalues $\widetilde{\lambda}_j$ of the symmetric matrix $\widetilde{S}$, $j = 1, \ldots, p$, $j \neq k$. In the case where several observations are deleted this author suggests a fairly lengthy procedure assuming that these observations are removed one after the other, so that we have a series of rank one perturbations.

New Developments

Some Relations Between the Approximations

In his work, Bénasséni (1987) only considers the approximations of order one $\widetilde{\lambda}_k^{(1)}$ for the eigenvalues of $\widetilde{S}$ and provides no comparison of $\widetilde{\lambda}_k^{(2)}$ with approximations based on

Rayleigh quotients. However, it is easy to derive the following simple relations. First note that $q_k^{(0)}$ can be written as:

$$q_k^{(0)} = \lambda_k + \left(\frac{r}{n-r}\right)\left(\lambda_k - \frac{1}{r}\sum_{i \in I}\alpha_{ki}^2\right) - \left(\frac{r}{n-r}\right)^2 \alpha_{kI}^2 \quad (11)$$

using (1). Therefore we see that we have always $\widetilde{\lambda}_k^{(1)} \geq q_k^{(0)}$. Furthermore, a simple comparison of (11) with (5) shows that:

$$\widetilde{\lambda}_k^{(2)} - q_k^{(0)} = -\left(\frac{r}{n-r}\right)^2 \sum_{j \neq k} \frac{1}{\lambda_j - \lambda_k}\left(\frac{1}{r}\sum_{i \in I}\alpha_{ki}\alpha_{ji}\right)^2.$$

In particular, when focusing on the largest eigenvalue which plays a central role in PCA, this difference is non negative so that we have $\widetilde{\lambda}_1^{(2)} \geq q_1^{(0)}$. In a similar way, when considering the smallest eigenvalue, we get $\widetilde{\lambda}_p^{(2)} \leq q_p^{(0)}$.

We omit the derivation of $q_k^{(1)}$ which is tedious and leads to a formula too complicated to be interpreted. However, it should be noted that this approximation involves terms up to the order 4 in $\epsilon = \frac{r}{n-r}$.

Improved Inequalities in Error Analysis

In practice, it is of crucial importance to have bounds as close as possible to the true values of $|\widetilde{\lambda}_k - q_k^{(m)}|$ and $\cos\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(m)}\right)$ in Inequalities (8) and (9). This will allow for an evaluation as precise as possible of the real accuracy of the approximations without having to recompute the perturbed analysis. For this purpose, we introduce below refined inequalities in the study of covariance matrices in order to get an improved error analysis

Indeed, Inequalities (7), (8) and (9) introduced in the Subsection Error Analysis can be improved using error analysis developed in Chatelin (2012, pp. 180–184). More precisely, it is easily derived from Corollary 4.6.4 in this reference that:

$$|\widetilde{\lambda}_k - q_k^{(m)}| \leq \frac{\left(\eta_k^{(m)}\right)^2}{c_k^{(m)}} \quad (12)$$

and

$$\sin\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(m)}\right) \leq \frac{\eta_k^{(m)}}{c_k^{(m)}} \quad (13)$$

under the condition:

$$\eta_k^{(m)} < c_k^{(m)}. \tag{14}$$

It is obvious that (12) is more accurate than (8). A similar remark holds for (13) wich improves (9). This last point is easily checked by converting (13) into

$$cos^2\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(m)}\right) \geq 1 - \left(\eta_k^{(m)} / c_k^{(m)}\right)^2. \tag{15}$$

Then letting $a = \left(\eta_k^{(m)} / c_k^{(m)}\right)^2$, $A = 1 - a$ and $B = \left[1 - (a/2)(1 + a)\right]^2$, Inequality (9) becomes $cos^2\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(m)}\right) \geq B$ so that we have simply to prove that $A \geq B$. Developing B, we get $B = A + (a^2/4)(a^2 + 2a - 3)$. Thus we have that $A \geq B$ if and only if the polynomial $a^2 + 2a - 3$ is non-positive. This is the case when a belongs to $[-3, 1]$. Therefore the result follows since $0 \leq a < 1$ from (14).

Furthermore, when dealing with the eigenvector associated to the largest eigenvalue, Chatelin (2012, p. 204) points out that Inequality (13) can also be refined into the following tangent based inequality:

$$\tan\left(\widetilde{\boldsymbol{\phi}}_1, \widetilde{\boldsymbol{\phi}}_1^{(m)}\right) \leq \frac{\eta_1^{(m)}}{c_1^{(m)}} \tag{16}$$

since this eigenvalue is assumed to be simple. More precisely, letting α denote the angle between the two vectors $\widetilde{\boldsymbol{\phi}}_1$ and $\widetilde{\boldsymbol{\phi}}_1^{(m)}$, we have $\alpha \leq \arctan\left(\eta_1^{(m)} / c_1^{(m)}\right) \leq \arcsin\left(\eta_1^{(m)} / c_1^{(m)}\right)$ since $0 \leq \arctan x \leq \arcsin x$ for all $x \in [0, 1]$. Thus we obtain a better approximation of α when using the function arctan rather than the function arcsin showing that (16) improves (13).

Improved Value for the Parameter $c_k^{(m)}$

The sharpness of the bounds in Inequalities (12), (13) and (16) depends on the value of the parameter $c_k^{(m)}$. The larger $c_k^{(m)}$ is, the sharper are these inequalities. In order to get a suitable value for this parameter, some other results of Chatelin (2012, pp. 180–181) or of Wilkinson (1988, pp. 174–176) turn out to be also of specific interest since they often improve significantly the value of the parameter $c_k^{(m)}$ obtained through the Courant-Fischer theorem in Bénasséni (1987). More precisely, from these references we know that there is at least one eigenvalue of $\widetilde{\mathrm{S}}$ in each of the intervals defined for $j = 1, \ldots, p$ by $\left[b_j^{(m)}, B_j^{(m)}\right] = \left[q_j^{(m)} - \eta_j^{(m)}, q_j^{(m)} + \eta_j^{(m)}\right]$ which are often referred to as Krylov-Weinstein intervals. When the interval $\left[b_k^{(m)}, B_k^{(m)}\right]$ is isolated from the $p - 1$ other ones, we know that it contains precisely one eigenvalue. Then, assuming that the Rayleigh quotients satisfy $q_1^{(m)} > q_2^{(m)} > \ldots > q_p^{(m)}$ (after having been reordered if necessary), a value for $c_k^{(m)}$ can be easily derived as $c_k^{(m)} = \min\left(b_k^{(m)} - B_{k+1}^{(m)}, b_{k-1}^{(m)} - B_k^{(m)}\right)$ if $k \in \{2, \ldots, p-1\}$, $c_1^{(m)} = b_1^{(m)} - B_2^{(m)}$ if $k = 1$ and $c_p^{(m)} = b_{p-1}^{(m)} - B_p^{(m)}$ if $k = p$. Finally, note that these Krylov-Weinstein intervals proved to be also useful in studying the effects of some small errors in the data table itself as emphasized by the work of Bénasséni (1988, pp. 303–310)

It should be pointed out that for very close eigenvalues, giving a value to this parameter can be a real issue. However, once a value satisfying (14) is obtained, we know from (12) that, for $k = 1, \ldots, p$, the eigenvalues $\widetilde{\lambda}_k$ lie in the intervals $\left[q_k^{(m)} - (\eta_k^{(m)})^2 / c_k^{(m)}, q_k^{(m)} + (\eta_k^{(m)})^2 / c_k^{(m)}\right]$ which are sharper than the Krylov-Weinstein intervals as soon as (14) holds. These new intervals can be used to obtain a larger value for the parameter $c_k^{(m)}$, thus improving (12), (13) and (16). This process could be iterated, but no significative improvment is generally observed.

Finally, it was noted in (10) that we have a rank one perturbation when deleting a single observation. Several inequalities and relations for this restricted rank perturbation are given in Bénasséni (1987), Hadi and Nyquist (1993), and Wang and Nyquist (1991). The reader is also referred to more recent works by Bénasséni (2011), Cheng et al. (2014), and Ipsen and Nadler (2009), who suggest new bounds for the perturbed eigenvalues. Although Krylov-Weinstein intervals generally provide a satisfying value for the parameter $c_k^{(m)}$, these recent works can also be interesting in the determination of the largest possible constant $c_k^{(m)}$ in order to make Inequalities (12), (13) and (16) sharper.

Numerical Study

The numerical illustration of the results is based on the soil composition data in Kendall (1975) which have already been used in several works including, among others, Bénasséni (2018), Critchley (1985), Enguix-González et al. (2005), Enguix-González et al. (2012), Tanaka (1988), Wang and Liski (1993), or Wang and Nyquist (1991) for sensitivity study of covariance based principal component analysis. The data table consists of 20 observations measured on 4 variables. We have the following four eigenvalues for the corresponding covariance matrix: $\lambda_1 = 82.30827$, $\lambda_2 = 6.73891$, $\lambda_3 = 0.44783$, $\lambda_4 = 0.24552$. In the first subsection, we study the effect of deleting each of the 20 observations on the two largest eigenvalues (which account for more than 99% of the total variation in principal component analysis) and on their corresponding eigenvectors. In the following subsection, we study the effects of deleting subsets of two observations on the largest eigenvalue. These subsets are those considered for the numerical study in Wang and Liski (1993). Finally, a short illustration of the potential effects of removing three observations is given in the last subsection.

Approximations When Deleting One Observation

Table 1 provides for each subset $I = \{i\}$, the perturbed eigenvalue $\widetilde{\lambda}_1$, its order one and two approximations $\widetilde{\lambda}_1^{(1)}$ and $\widetilde{\lambda}_1^{(2)}$, the Rayleigh quotients $q_1^{(0)}$ ans $q_1^{(1)}$ and the differences between each approximation and the true perturbed eigenvalue $\widetilde{\lambda}_1$. In the last two columns we find the bounds to $|\widetilde{\lambda}_1 - q_1^{(0)}|$ and $|\widetilde{\lambda}_1 - q_1^{(1)}|$ given by Inequality (12).

The first comment regarding the results in this table is that $\widetilde{\lambda}_1^{(1)}$ is by far the less accurate approximation in all cases. In contrast $q_1^{(1)}$ always provides extremely sharp approximations since it deviates from $\widetilde{\lambda}_1$ by 1.76×10^{-3} in the worst case (when deleting Observation 4) and that the error is only 2.13×10^{-13} when deleting Observation 16. It should be noted that $\widetilde{\lambda}_1^{(2)}$ also provides fairly satisfying approximations although clearly less accurate than $q_1^{(1)}$. The Rayleigh quotient $q_1^{(0)}$ is outperformed by $\widetilde{\lambda}_1^{(2)}$ but remains significantly sharper than $\widetilde{\lambda}_1^{(1)}$. Furthermore, it is worth pointing out that $\widetilde{\lambda}_1^{(1)}$ always overestimates the perturbed eigenvalue while the other three estimations slightly underestimate it. Note also that the results agree with inequalities $\widetilde{\lambda}_1^{(1)} \geq q_1^{(0)}$ and $\widetilde{\lambda}_1^{(2)} \geq q_1^{(0)}$ given in the Subsection Some Relations Between the Approximations.

Second, Inequality (12) provides bounds sufficiently close to $|\widetilde{\lambda}_1 - q_1^{(0)}|$ and $|\widetilde{\lambda}_1 - q_1^{(1)}|$ to evaluate correctly the accuracy of Rayleigh quotients as approximations to $\widetilde{\lambda}_1$ without having to recompute the perturbed analysis.

Third, it is easily seen from (10) that the maximum value for the perturbed eigenvalue is obtained when $\mathbf{x}_i = \bar{\mathbf{x}}$ with $\widetilde{\lambda}_1 = (20\,\lambda_1)/19 = 86.64028$. We have the highest perturbed eigenvalue when deleting Observations Number 3, 5, 10, 12, 14, 16, 20 which are fairly close to $\bar{\mathbf{x}}$ and in these cases we get the sharper approximations to $\widetilde{\lambda}_1$.

Focusing now on the second largest eigenvalue, Table 2 provides results similar to those of Table 1. It turns out that $\widetilde{\lambda}_2^{(1)}$ is the less accurate approximation to $\widetilde{\lambda}_2$. Except when deleting Observation 13, the Rayleigh quotient $q_2^{(1)}$ again provides the best approximation with a very good accuracy since in the worst case corresponding to this observation we have $\widetilde{\lambda}_2 - q_2^{(1)} = 7.69 \times 10^{-4}$. For this observation, $\widetilde{\lambda}_2^{(2)}$ is slightly better but less accurate in all the other cases while performing fairly well in general. The Rayleigh quotient $q_2^{(0)}$ performs in a similar way as $q_1^{(0)}$ in Table 1.

It should be noted that again $\widetilde{\lambda}_2^{(1)}$ always overestimates the perturbed eigenvalue but in contrast to Table 1, the other three approximations can as well slightly underestimate or overestimate $\widetilde{\lambda}_2$.

Table 1

Approximations and Error Analysis for the Largest Eigenvalue With r = 1

I	$\widetilde{\lambda}_1$	$\widetilde{\lambda}_1^{(1)}$	$\widetilde{\lambda}_1^{(2)}$	$q_1^{(0)}$	$q_1^{(1)}$	$\widetilde{\lambda}_1 - \widetilde{\lambda}_1^{(1)}$	$\widetilde{\lambda}_1 - \widetilde{\lambda}_1^{(2)}$	$\widetilde{\lambda}_1 - q_1^{(0)}$	$\widetilde{\lambda}_1 - q_1^{(1)}$	$\lvert\widetilde{\lambda}_1 - q_1^{(0)}\rvert \le$	$\lvert\widetilde{\lambda}_1 - q_1^{(1)}\rvert \le$
1	81.59976	81.8333	81.59769	81.58031	81.59969	-0.23355	0.00207	0.01945	6.87×10^{-5}	0.01988	6.91×10^{-5}
2	77.18681	77.64932	77.18511	77.17611	77.18667	-0.46251	0.00170	0.01070	1.42×10^{-4}	0.01115	1.46×10^{-4}
3	86.52309	86.52844	86.52306	86.52255	86.52309	-0.00535	2.53×10^{-5}	5.38×10^{-4}	5.20×10^{-9}	5.40×10^{-4}	5.22×10^{-9}
4	77.34975	77.64477	77.32399	77.17133	77.34799	-0.29503	0.02576	0.17842	1.76×10^{-3}	0.18922	1.78×10^{-3}
5	86.57878	86.58112	86.57875	86.57801	86.57878	-0.00234	2.99×10^{-5}	7.71×10^{-4}	1.06×10^{-7}	7.74×10^{-4}	1.07×10^{-7}
6	76.85039	77.33877	76.8502	76.84922	76.85037	-0.48838	1.86×10^{-4}	1.17×10^{-3}	1.53×10^{-5}	1.28×10^{-3}	1.66×10^{-5}
7	79.95562	80.27544	79.95368	79.94044	79.95552	-0.31981	0.00194	0.01518	1.02×10^{-4}	0.01544	1.02×10^{-4}
8	74.71505	75.29919	74.71264	74.70229	74.71477	-0.58414	0.00241	0.01276	2.72×10^{-4}	0.01326	2.79×10^{-4}
9	74.45165	75.05167	74.4498	74.44174	74.45144	-0.60002	0.00185	9.91×10^{-3}	2.04×10^{-4}	0.01073	2.19×10^{-4}
10	86.02449	86.05404	86.02442	86.02318	86.02449	-0.02955	7.21×10^{-5}	1.30×10^{-3}	3.99×10^{-8}	1.33×10^{-3}	4.04×10^{-8}
11	85.32593	85.38314	85.3254	85.31697	85.32593	-0.05721	5.31×10^{-4}	0.001782881	8.62×10^{-7}	9.06×10^{-3}	8.63×10^{-7}
12	86.58382	86.58643	86.58381	86.58359	86.58382	-0.00261	1.06×10^{-5}	2.26×10^{-4}	2.48×10^{-9}	2.32×10^{-4}	2.53×10^{-9}
13	84.13095	84.20755	84.12781	84.07951	84.13094	-0.07660	0.00314	0.05143	6.97×10^{-6}	0.05293	6.99×10^{-6}
14	86.52446	86.52855	86.52439	86.52267	86.52446	-0.00410	6.56×10^{-5}	1.78×10^{-3}	3.46×10^{-7}	1.82×10^{-3}	3.50×10^{-7}
15	82.72790	82.91768	82.72732	82.72175	82.72788	-0.18978	5.84×10^{-4}	6.15×10^{-3}	1.37×10^{-5}	6.26×10^{-3}	1.39×10^{-5}
16	86.63917	86.63923	86.63917	86.63917	86.63917	-5.48×10^{-5}	3.36×10^{-8}	6.79×10^{-7}	2.13×10^{-13}	7.21×10^{-7}	2.59×10^{-13}
17	83.27902	83.42004	83.27668	83.25055	83.27899	-0.14101	0.00234	0.02847	3.27×10^{-5}	0.02925	3.29×10^{-5}
18	80.80479	81.08976	80.80396	80.79763	80.80476	-0.28497	8.28×10^{-4}	7.16×10^{-3}	3.42×10^{-5}	7.57×10^{-3}	3.59×10^{-5}
19	78.33933	78.75217	78.33900	78.33701	78.33931	-0.41283	3.31×10^{-4}	2.33×10^{-3}	2.18×10^{-5}	2.53×10^{-3}	2.37×10^{-5}
20	86.37150	86.38477	86.37149	86.37133	86.37150	-0.01328	9.05×10^{-6}	1.73×10^{-4}	1.09×10^{-9}	1.86×10^{-4}	1.17×10^{-9}

Table 2

Approximations and Error Analysis for the Second Largest Eigenvalue with r = 1

I	$\widetilde{\lambda}_2$	$\widetilde{\lambda}_2^{(1)}$	$\widetilde{\lambda}_2^{(2)}$	$q_2^{(0)}$	$q_2^{(1)}$	$\widetilde{\lambda}_2 - \widetilde{\lambda}_2^{(1)}$	$\widetilde{\lambda}_2 - \widetilde{\lambda}_2^{(2)}$	$\widetilde{\lambda}_2 - q_2^{(0)}$	$\widetilde{\lambda}_2 - q_2^{(1)}$	$\lvert\widetilde{\lambda}_2 - q_2^{(0)}\rvert \leq$	$\lvert\widetilde{\lambda}_2 - q_2^{(1)}\rvert \leq$
1	6.80760	6.83824	6.80935	6.82480	6.80764	-0.03065	-0.00175	-0.01721	-4.90×10^{-5}	0.22557	7.06×10^{-4}
2	7.03262	7.04231	7.03372	7.03961	7.03271	-0.00969	-0.00110	-0.00699	-9.02×10^{-5}	0.08423	1.02×10^{-3}
3	6.73724	6.75509	6.73721	6.73727	6.73724	-0.01785	2.51×10^{-5}	-3.39×10^{-5}	1.36×10^{-6}	7.24×10^{-3}	1.53×10^{-6}
4	5.66955	5.87923	5.68474	5.81532	5.66891	-0.20968	-0.01519	-0.14577	6.36×10^{-4}	2.52388	0.02060
5	6.10058	6.15035	6.10049	6.10071	6.10057	-0.04977	9.24×10^{-5}	-1.22×10^{-4}	1.36×10^{-5}	0.01169	1.59×10^{-5}
6	7.09264	7.0928	7.09266	7.09276	7.09264	-1.56×10^{-4}	-1.90×10^{-5}	-1.15×10^{-4}	-1.71×10^{-6}	1.23×10^{-3}	1.82×10^{-5}
7	6.91733	6.94011	6.91920	6.93203	6.91743	-0.02277	-0.00187	-0.01470	-9.66×10^{-5}	0.17207	1.12×10^{-3}
8	7.03111	7.04295	7.03284	7.04029	7.03131	-0.01184	-0.00173	-0.00917	-1.96×10^{-4}	0.09986	2.08×10^{-3}
9	7.07835	7.08123	7.07877	7.080583	7.07840	-0.00288	-4.22×10^{-4}	-2.23×10^{-3}	-4.77×10^{-5}	0.02615	5.16×10^{-4}
10	6.95776	6.96488	6.95776	6.95811	6.95776	-0.00712	-3.17×10^{-6}	-3.48×10^{-4}	3.59×10^{-7}	0.01375	3.83×10^{-7}
11	6.55595	6.59092	6.55642	6.56447	6.55595	-0.03497	-4.69×10^{-4}	-0.00851	2.69×10^{-6}	0.11689	1.33×10^{-5}
12	6.85444	6.86361	6.85422	6.85150	6.85444	-0.00917	2.21×10^{-4}	0.00293	1.79×10^{-6}	0.00536	2.30×10^{-6}
13	5.50884	5.62489	5.50826	5.54759	5.50807	-0.11605	5.83×10^{-4}	-0.03875	7.69×10^{-4}	0.84790	8.23×10^{-4}
14	6.04911	6.07404	6.04375	6.02038	6.04857	-0.02493	0.00536	0.02873	5.42×10^{-4}	0.05689	5.82×10^{-4}
15	6.98255	6.99306	6.98304	6.98777	6.98256	-0.01051	-4.84×10^{-4}	-0.00522	-1.04×10^{-5}	0.06610	1.34×10^{-4}
16	7.08039	7.08097	7.08038	7.08031	7.08039	-5.88×10^{-4}	3.51×10^{-6}	7.55×10^{-5}	1.05×10^{-9}	7.94×10^{-5}	1.15×10^{-9}
17	6.48050	6.53032	6.48147	6.50067	6.48043	-0.04981	-9.70×10^{-4}	-0.02017	7.06×10^{-5}	0.3604578	3.20×10^{-4}
18	7.04966	7.05459	7.04998	7.05253	7.04967	-0.00492	-3.18×10^{-4}	-2.87×10^{-3}	-1.18×10^{-5}	0.03878	1.56×10^{-4}
19	7.09124	7.09159	7.09128	7.09148	7.09125	-3.43×10^{-4}	-3.48×10^{-5}	-2.38×10^{-4}	-2.45×10^{-6}	2.68×10^{-3}	2.73×10^{-5}
20	7.08664	7.08696	7.08664	7.08662	7.08664	-3.23×10^{-4}	1.23×10^{-6}	2.54×10^{-5}	9.25×10^{-12}	3.36×10^{-4}	4.31×10^{-9}

Table 3

Approximations and Error Analysis for the Eigenvectors Associated to the Two Largest Eigenvalues with r = 1

I	$\sin(\tilde{\phi}_1, \phi_1)$	$\sin(\tilde{\phi}_1, \tilde{\phi}_1^{(1)})$	$\sin(\tilde{\phi}_1, \phi_1) \leq$	$\sin(\tilde{\phi}_1, \tilde{\phi}_1^{(1)}) \leq$	$\sin(\tilde{\phi}_2, \phi_2)$	$\sin(\tilde{\phi}_2, \tilde{\phi}_2^{(1)})$	$\sin(\tilde{\phi}_2, \phi_2) \leq$	$\sin(\tilde{\phi}_2, \tilde{\phi}_2^{(1)}) \leq$
1	0.01609	9.56×10^{-4}	0.01644	9.62×10^{-4}	0.01988	0.00153	0.19303	0.01056
2	0.01218	1.40×10^{-3}	0.01267	1.44×10^{-3}	0.01213	0.00142	0.11768	0.01250
3	0.00259	8.07×10^{-6}	0.00261	8.09×10^{-6}	0.00912	4.62×10^{-4}	0.03408	4.94×10^{-4}
4	0.04974	4.94×10^{-3}	0.05280	4.99×10^{-3}	0.08114	0.01954	0.72248	0.06298
5	0.00309	3.64×10^{-5}	0.00311	3.64×10^{-5}	0.01097	0.00154	0.04563	0.00168
6	0.00393	4.50×10^{-4}	0.00428	4.88×-10^{-3}	0.01491	0.00125	0.16533	0.01317
8	0.01356	1.98×10^{-3}	0.01408	2.03×10^{-3}	0.01310	0.00194	0.12507	0.01782
9	0.01169	1.68×10^{-3}	0.01266	1.80×10^{-3}	0.00725	0.00107	0.06553	0.00883
10	0.00403	2.23×10^{-5}	0.00411	2.26×10^{-5}	0.01092	2.32×10^{-4}	0.04628	2.43×10^{-4}
11	0.01066	1.05×10^{-4}	0.01078	1.05×10^{-4}	0.01321	7.59×10^{-4}	0.13947	0.00148
12	0.00167	5.52×10^{-6}	0.00171	5.63×10^{-6}	0.02195	5.32×10^{-4}	0.02915	5.96×10^{-4}
13	0.02555	2.98×10^{-4}	0.02630	2.98×10^{-4}	0.05398	0.01230	0.42319	0.01285
14	0.00468	6.52×10^{-5}	0.00477	6.60×10^{-5}	0.07252	0.00970	0.10386	0.01011
15	0.00897	4.24×10^{-4}	0.00913	4.28×10^{-4}	0.01053	5.52×10^{-4}	0.10247	0.00454
16	8.97×10^{-5}	5.16×10^{-8}	9.52×10^{-5}	5.70×10^{-8}	0.00335	1.25×10^{-5}	0.00347	1.32×10^{-5}
17	0.01918	6.50×10^{-4}	0.01971	6.55×10^{-4}	0.03624	0.00383	0.25215	0.00729
18	0.00962	6.66×10^{-4}	0.01017	6.97×10^{-4}	0.00999	6.86×10^{-4}	0.07882	0.00486
19	0.00549	5.31×10^{-4}	0.00597	5.76×10^{-4}	0.00213	2.16×10^{-4}	0.02021	0.00203
20	0.00142	3.58×10^{-6}	0.00153	3.84×10^{-6}	0.00274	7.34×10^{-6}	0.00715	2.55×10^{-5}

Another difference with Table 1 is that bounds provided to $|\widetilde{\lambda}_2 - q_2^{(0)}|$ and $|\widetilde{\lambda}_2 - q_2^{(1)}|$ by Inequality (12) are not so close to these quantities as they were previously. For a part, this can be explained by the fact that we have a smaller value for $c_2^{(0)}$ and $c_2^{(1)}$ than for $c_1^{(0)}$ and $c_1^{(1)}$ when considering the largest eigenvalue. Indeed, for $m = 0, 1$ the gap $|\widetilde{\lambda}_3 - q_2^{(m)}|$ is smaller than the gap $|\widetilde{\lambda}_2 - q_1^{(m)}|$. However, if we except the case of Observation 4, we know from this bound that $|\widetilde{\lambda}_2 - q_2^{(1)}|$ never exceeds 1.02×10^{-3} and this is sufficient for practical interpretation.

Results for the eigenvectors corresponding to the two largest eigenvalues are given in Table 3 which provides the sines $\sin(\widetilde{\boldsymbol{\phi}}_k, \boldsymbol{\phi}_k)$ and $\sin\left(\widetilde{\boldsymbol{\phi}}_k, \widetilde{\boldsymbol{\phi}}_k^{(1)}\right)$ for $k = 1, 2$ and their bounds provided by Inequality (13).

First, we consider the eigenvector associated to the largest eigenvalue. It could be noted that the maximum value of the sine bewteen the unperturbed and perturbed eigenvector is obtained when deleting Observation 4. This value corresponds to an angle of 2.85°. In this case, as in all the other ones, the order one approximation performs fairly well since its sine with the perturbed eigenvector is equal to only 0.00494 which corresponds to an angle of 0.28°. Furthermore bounds provided by Inequality (13) are always extremely close to the exact value of $\sin\left(\widetilde{\boldsymbol{\phi}}_1, \widetilde{\boldsymbol{\phi}}_1^{(1)}\right)$.

Focusing now on the eigenvector corresponding to the second largest eigenvalue, the maximum of $\sin(\widetilde{\boldsymbol{\phi}}_2, \boldsymbol{\phi}_2)$ is again obtained when deleting Observation 4 with the value of 0.08114. Even in this case, the order one approximation is fairly close to the perturbed eigenvector since $\sin\left(\widetilde{\boldsymbol{\phi}}_2, \widetilde{\boldsymbol{\phi}}_2^{(1)}\right) = 0.01954$. In contrast to the previous eigenvector, it is worth pointing out that bounds provided by Inequality (13) are not always sufficiently close to the true values of the sine to give an exact account of the accuracy for these approximations.

Finally, since the angles between the eigenvectors studied in the table are always very close to zero, we do not provide the tangent of these angles which only deviates from the sine by an extremely small amount.

Approximations When Deleting Subsets of Two Observations

Now, we study approximations to the perturbed largest eigenvalue ant its corresponding eigenvector when deleting the subsets of two observations considered for the numerical illustration in Wang and Liski (1993). Results similar to those of the previous subsection are provided in Tables 4 and 5.

Table 4

Approximations and Error Analysis for the Largest Eigenvalue with r = 2

I	$\widetilde{\lambda}_1$	$\widetilde{\lambda}_1^{(1)}$	$\widetilde{\lambda}_1^{(2)}$	$q_1^{(0)}$	$q_1^{(1)}$	$\widetilde{\lambda}_1 - \widetilde{\lambda}_1^{(1)}$	$\widetilde{\lambda}_1 - \widetilde{\lambda}_1^{(2)}$	$\widetilde{\lambda}_1 - q_1^{(0)}$	$\widetilde{\lambda}_1 - q_1^{(1)}$	$\lvert\widetilde{\lambda}_1 - q_1^{(0)}\rvert \leq$	$\lvert\widetilde{\lambda}_1 - q_1^{(1)}\rvert \leq$
$\{8, 9\}$	64.59310	67.25005	64.58126	64.56084	64.59026	-2.65695	0.01184	0.03226	0.00284	0.03616	0.00316
$\{4, 9\}$	67.51718	69.72595	67.4481	67.32138	67.50241	-2.20877	0.06908	0.19580	0.01476	0.20842	0.01501
$\{4, 8\}$	67.98219	69.98722	67.85527	67.61002	67.95618	-2.00502	0.12692	0.37218	0.02602	0.40664	0.02660
$\{6, 9\}$	69.40391	69.40294	69.40242	69.39558	69.40366	9.72×10^{-4}	0.00149	0.00834	2.52×10^{-4}	0.00892	2.66×10^{-4}
$\{2, 9\}$	69.74526	69.73074	69.74115	69.72109	69.74455	0.01451	0.00410	0.02417	7.04×10^{-4}	0.02512	7.19×10^{-4}
$\{2, 8\}$	69.98911	69.99202	69.98832	69.98402	69.98899	-0.00291	7.86×10^{-4}	0.00509	1.21×10^{-4}	0.00565	1.33×10^{-4}
$\{2, 6\}$	70.01929	72.14491	70.01294	69.99965	70.01808	-2.12561	0.00636	0.01964	0.00121	0.02072	0.00126
$\{3, 8\}$	78.83587	79.36442	78.83242	78.82489	78.83542	-0.52854	0.00345	0.01099	4.54×10^{-4}	0.01144	4.59×10^{-4}
$\{2, 4\}$	72.60356	72.46791	72.58774	72.46791	72.60172	0.13565	0.01582	0.13565	0.00184	0.14414	0.00188
$\{4, 7\}$	75.53263	75.23992	75.50596	75.22661	75.53053	0.29271	0.02667	0.30602	0.00210	0.33028	0.00212

Table 5

Approximations and Error Analysis for the Eigenvector Associated to the Largest Eigenvalue with r = 2

$\boldsymbol{I}$	$\sin(\tilde{\boldsymbol{\phi}}_1, \boldsymbol{\phi}_1)$	$\sin(\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_1^{(1)})$	$\sin(\tilde{\boldsymbol{\phi}}_1, \boldsymbol{\phi}_1) \leq$	$\sin(\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_1^{(1)}) \leq$	$\boldsymbol{I}$	$\sin(\tilde{\boldsymbol{\phi}}_1, \boldsymbol{\phi}_1)$	$\sin(\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_1^{(1)})$	$\sin(\tilde{\boldsymbol{\phi}}_1, \boldsymbol{\phi}_1) \leq$	$\sin(\tilde{\boldsymbol{\phi}}_1, \tilde{\boldsymbol{\phi}}_1^{(1)}) \leq$
$\{8, 9\}$	0.02254	0.00667	0.02525	0.00745	$\{2, 8\}$	0.00856	0.00132	0.00951	0.00146
$\{4, 9\}$	0.05637	0.01548	0.06010	0.01574	$\{2, 6\}$	0.01740	0.00432	0.01833	0.00449
$\{4, 8\}$	0.07724	0.02042	0.08464	0.02089	$\{3, 8\}$	0.01220	0.00251	0.01269	0.00253
$\{6, 9\}$	0.01128	0.00197	0.01205	0.00207	$\{2, 4\}$	0.04477	0.00522	0.04759	0.00532
$\{2, 9\}$	0.01950	0.00333	0.02026	0.00340	$\{4, 7\}$	0.06609	0.00548	0.07148	0.00553

We note that deleting subsets of two observations can result in larger variations of the eigenvalue of interest than when deleting single observations. Indeed, the perturbed eigenvalue is lower than 70 in the first six lines of Table 4.

Furthermore for all the subsets I studied in this table we have a decrease of the eigenvalue, while we note that this eigenvalue is increased in several cases in Table 1. Despite these significant variations of the eigenvalue, we see that the Rayleigh quotient $q_1^{(1)}$ always provides a very accurate approximation to $\widetilde{\lambda}_1$ since the maximum gap $\left|\widetilde{\lambda}_1 - q_1^{(1)}\right| = 0.02602$ observed for $I = \{4, 8\}$ remains fairly moderate. It should also be noted that $q_1^{(1)}$ always performs better than $\widetilde{\lambda}_1^{(2)}$. This point is fairly well illustrated considering again the case of $I = \{4, 8\}$ for which we have $\widetilde{\lambda}_1 - \widetilde{\lambda}_1^{(2)} = 0.12692$. The Rayleigh quotient $q_1^{(0)}$ provides less accurate approximations than $\widetilde{\lambda}_1^{(2)}$ but should generally be preferred to $\widetilde{\lambda}_1^{(1)}$ if we except some cases. Finally, it is worth pointing out that bounds to $\left|\widetilde{\lambda}_1 - q_1^{(0)}\right|$ and $\left|\widetilde{\lambda}_1 - q_1^{(1)}\right|$ provided by Inequality (12) are always very close to the true values of these two differences thus avoiding to recompute the perturbed analysis.

Turning now to the sine values in Table 5 , we note the largest variations of the eigenvector when deleting the subsets $I = \{4, 7\}$, $I = \{4, 8\}$ and $I = \{4, 9\}$. However, the order one approximation $\widetilde{\boldsymbol{\phi}}_1^{(1)}$ remains fairly satisfying in all the cases since the maximum value of $\sin\left(\widetilde{\boldsymbol{\phi}}_1, \widetilde{\boldsymbol{\phi}}_1^{(1)}\right)$ obtained when deleting the subset $I = \{4, 8\}$ does not exceed 0.02042 which corresponds to an angle of only 1.17°.

Approximations When Deleting Subsets of Three Observations

While the accuracy of $q_1^{(1)}$ is fairly good for the removal of two observations despite the small sample size, one might think that results would probably be much worse when removing three observations and that their performance would deteriorate very quickly as $\epsilon = \frac{r}{n-r}$ increases. In order to investigate this point, we decided to perform a very brief study of the performance of the approximations for the largest eigenvalue when removing some sets of three observations. Note that in this case we have: $\epsilon = \frac{3}{17} = 0.1765$ instead of $\epsilon = \frac{2}{18} = 0.1111$ when deleting only two observations.

First, when removing the three Observations 3, 4 and 7, the largest eigenvalue initially equal to $\lambda_1 = 82.30827$ is moderately decreased to $\widetilde{\lambda}_1 = 79.83362$. The approximations to $\widetilde{\lambda}_1$ are then the following: $\widetilde{\lambda}_1^{(1)} = 79.5408$, $\widetilde{\lambda}_1^{(2)} = 79.82602$, $q_1^{(0)} = 79.53947$ and $q_1^{(1)} = 79.83346$. We note the excellent accuracy of $q_1^{(1)}$ since the error $\widetilde{\lambda}_1 - q_1^{(1)} = 0.00016$ is negligible.

Second, when removing the three Observations 2, 4 and 6, the largest eigenvalue is now significantly decreased to $\widetilde{\lambda}_1 = 65.86825$. The approximations to $\widetilde{\lambda}_1$ are then the following: $\widetilde{\lambda}_1^{(1)} = 66.33492$, $\widetilde{\lambda}_1^{(2)} = 65.85841$, $q_1^{(0)} = 65.7237$ and $q_1^{(1)} = 65.86529$. We still have a satisfying accuracy of $q_1^{(1)}$ since the error $\widetilde{\lambda}_1 - q_1^{(1)}$ is equal to 0.00295.

Finally, when removing the very influential set $\{4, 8, 9\}$, the largest eigenvalue is drastically decreased to $\widetilde{\lambda}_1 = 55.36930$ and we get the following approximations: $\widetilde{\lambda}_1^{(1)} = 61, 15213$, $\widetilde{\lambda}_1^{(2)} = 55.10147$, $q_1^{(0)} = 54.87525$ and $q_1^{(1)} = 55.26709$. Therefore while the largest eigenvalue is decreased by 26.939 from its initial value, the error $\widetilde{\lambda}_1 - q_1^{(1)} = 0.10221$ remains fairly moderate while obviously larger than when considering removal of only two observations.

However, it would be necessary to perform additional studies with other data tables to get more reliable information on the relation between the value of ϵ and the accuracy of the approximations.

Concluding Remarks

The previous numerical study provides some indications on the sharpness of the various approximations in the paper when considering a particular data set. As a result, it may be useful to provide practitioners with some guidance on the choice of approximations for perturbed covariance matrices.

First, when focusing on eigenvalues, it should be noted that, in this example, approximations provided by $\widetilde{\lambda}_k^{(1)}$ should be avoided as well as the Rayleigh quotients $q_k^{(0)}$ which are not sufficiently accurate.

In contrast, Rayleigh quotients $q_k^{(1)}$ (based on the perturbed matrix $\widetilde{\mathrm{S}}$ and the approximations of order one $\widetilde{\boldsymbol{\phi}}_k^{(1)}$) seem to always provide reliable approximations to $\widetilde{\lambda}_k$. They must be preferred to the approximation of order two $\widetilde{\lambda}_k^{(2)}$ in the study of this data table. Furthermore, their accuracy can be evaluated in a precise way by Inequality (12) if the eigenvalue of interest is not too close to the other eigenvalues, as emphasized by results in Table 1 when considering the largest eigenvalue. This is a definite advantage over other approximations.

It should also be noted that these Rayleigh quotients perform fairly well even for values of $\epsilon = \frac{r}{n-r}$ which are not necessarily very close to zero. This point is made clear in the first lines of Table 4 where $\epsilon = \frac{2}{18} = 0.11$ and the eigenvalue $\widetilde{\lambda}_1$ is significantly decreased through the perturbation. The Subsection Approximations When Deleting Subsets of Three Observations, briefly dealing with the removal of three observations, tends to indicate that these Rayleigh quotients should still perform in a satisfying way even for higher values of ϵ. However, more in-depth studies would be necessary to get reliable conclusions on this point.

Second, when considering eigenvectors, we note a satisfying accuracy of approximations of order one $\widetilde{\boldsymbol{\phi}}_k^{(1)}$. Again, when the eigenvalue corresponding to the eigenvector of interest is sufficiently distant from the other ones, we have a correct evaluation of this accuracy by Inequality (13).

Funding: The authors have no funding to report.

Acknowledgments: The authors are grateful to the reviewers for their careful reading of the paper and their comments.

Competing Interests: The authors have declared that no competing interests exist.

References

Bénasséni, J. (1987). Perturbation des poids des unités statistiques et approximation en analyse en composantes principales. *RAIRO—Recherche opérationnelle/Operations Research, 21*(2), 175–198.

Bénasséni, J. (1988). Sensitivity of principal component to data perturbation (pp. 303–310). In E. Diday (Ed.), *Data analysis and informatics, V.* Elsevier.

Bénasséni, J. (2011). Lower bounds for the largest eigenvalue of a symmetric matrix under perturbations of rank one. *Linear and Multilinear Algebra, 59*(5), 565–569. https://doi.org/10.1080/03081081003709827

Bénasséni, J. (2018). A correction of approximations used in sensitivity study of principal component analysis. *Computational Statistics, 33*, 1939–1955. https://doi.org/10.1007/s00180-017-0790-7

Chatelin, F. (2012). *Eigenvalues of matrices.* Society for Industrial and Applied Mathematics.

Cheng, G., Song, Z., Yang, J., & Si, J. (2014). The bounds of the eigenvalues for rank-one modification of hermitian matrix. *Numerical Linear Algebra with Applications, 21*(1), 98–107. https://doi.org/10.1002/nla.1867

Critchley, F. (1985). Influence in principal component analysis. *Biometrika, 72*(3), 627–636. https://doi.org/10.1093/biomet/72.3.627

Enguix-González, A., Moreno-Rebollo, J. L., & Muñoz Pichardo, J. M. (2015). A better approximation of moments of the eigenvalues and eigenvectors of the sample covariance matrix. *Journal of Multivariate Analysis, 142*, 133–143. https://doi.org/10.1016/j.jmva.2015.08.002

Enguix-González, A., Muñoz Pichardo, J. M., Moreno-Rebollo, J. L., & Barranco-Chamorro, I. (2012). Using conditional bias in principal component analysis for the evaluation of joint influence on the eigenvalues of the covariance matrix. *Applied Mathematics and Computation, 218*(17), 8937–8950. https://doi.org/10.1016/j.amc.2012.02.054

Enguix-González, A., Muñoz Pichardo, J. M., Moreno-Rebollo, J. L., & Pino-Mejías, R. (2005). Influence analysis in principal component analysis through power-series expansions. *Communications in Statistics-Theory and Methods, 34*(9–10), 2025–2046. https://doi.org/10.1080/03610920500203505

Hadi, A., & Nyquist, H. (1993). Further theoretical results and a comparison between two methods for approximating eigenvalues of perturbed covariance matrices. *Statistics and Computing, 3*, 113–123. https://doi.org/10.1007/BF00147774

Ipsen, I. C. F., & Nadler, B. (2009). Refined perturbation bounds for eigenvalues of hermitian and non-hermitian matrices. *SIAM Journal on Matrix Analysis and Applications, 31*(1), 40–53. https://doi.org/10.1137/070682745

Jolliffe, I. T. (2002). *Principal component analysis* (2^{nd} ed.). Springer-Verlag.

Kendall, M. G. (1975). *Multivariate analysis.* Griffin.

Masioti, M., Li-Wai-Suen, C., Prendergast, L. A., & Shaker, A. (2023). A note on switching eigenvalues under small perturbations. *Communications in Statistics—Theory and Methods, 53*(20), 7311–7325. https://doi.org/10.1080/03610926.2023.2263114

Pack, P., Jolliffe, I. T., & Morgan, B. J. T. (1988). Influential observations in principal component analysis: A case-study. *Journal of Applied Statistics, 15*(1), 39–52. https://doi.org/10.1080/02664768800000004

Prendergast, L. A. (2008). A note on sensitivity of principal component subspaces and the efficient detection of influential observations in high dimensions. *Electronic Journal of Statistics, 2*, 454–467. https://doi.org/10.1214/08-EJS201

Prendergast, L. A., & Li-Wai-Suen, C. (2011). A new and practical influence measure for subsets of covariance matrix sample principal components with applications to high dimensional datasets. *Computational Statistics and Data Analysis, 55*(1), 752–764. https://doi.org/10.1016/j.csda.2010.06.022

Sibson, R. (1979). Studies in robustness of multidimensional scaling: perturbational analysis of classical scaling. *Journal of the Royal Statistical Society: Series B (Methodological), 41*(2), 217–229. https://doi.org/10.1111/j.2517-6161.1979.tb01076.x

Tanaka, Y. (1988). Sensitivity analysis in principal component analysis: Influence on the subspace spanned by principal components. *Communications in Statistics-Theory and Methods, 17*(9), 3157–3175. https://doi.org/10.1080/03610928808829796

Wang, S.-G., & Liski, E. P. (1993). Effects of observations on the eigensystem of a sample covariance matrix. *Journal of Statistical Planning and Inference, 36*, 215–226.

Wang, S.-G., & Nyquist, H. (1991). Effects on the eigenstructure of a data matrix when deleting an observation. *Computational Statistics and Data Analysis, 11*(2), 179–188. https://doi.org/10.1016/0167-9473(91)90068-D

Wilkinson, J. H. (1988). *The algebraic eigenvalue problem.* Oxford University Press.

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.

PsychOpen GOLD