Modern Stochastics: Theory and Applications 8 (4) (2021) 509–523
https://doi.org/10.15559/21-VMSTA191

Principal components analysis for mixtures with varying concentrations

Olena Sugakova[a], **Rostyslav Maiboroda**[a,*]

[a]*Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*

sugak@univ.kiev.ua (O. Sugakova), rostmaiboroda@knu.ua (R. Maiboroda)

Received: 14 August 2021, Revised: 30 October 2021, Accepted: 30 October 2021,
Published online: 12 November 2021

Abstract Principal Component Analysis (PCA) is a classical technique of dimension reduction for multivariate data. When the data are a mixture of subjects from different subpopulations one can be interested in PCA of some (or each) subpopulation separately. In this paper estimators are considered for PC directions and corresponding eigenvectors of subpopulations in the nonparametric model of mixture with varying concentrations. Consistency and asymptotic normality of obtained estimators are proved. These results allow one to construct confidence sets for the PC model parameters. Performance of such confidence intervals for the leading eigenvalues is investigated via simulations.

Keywords Finite mixture model, principal components, mixture with varying concentrations, nonparametric estimation, asymptotic normality, confidence interval, eigenvalue

2010 MSC 62J05, 62G20

1 Introduction

Principal components (PC) analysis is a standard technique of dimension reduction for multivariate data introduced by K. Pearson in 1901 and reinvented by H. Hotelling in the 1933 ([7], section 1.2). The first PC direction is the direction of the highest scattering of the data cloud and the first eigenvalue corresponding to it is the variance

*Corresponding author.

of the data projections on this direction. First two or three PC scores are usually used to visualize multidimensional data ([4], chapter 9).The orthogonal regression estimator for coefficients of a linear regression model is represented through the least PC direction (corresponding to the smallest eigenvalue, see (2.23) in [15]).

Classical PCA is developed for homogeneous samples. Real life statistical data is often a mixture of observations from different subpopulations with different distributions of observed variables. Finite mixture models (FMM) are developed to interpret such data. For parametric (normal) FMM the PCA provides a paradigm which allows one to describe and analyze multivariate data distribution of each subpopulation separately in straightforward and intuitive terms. Such an approach is used, e.g., in the `R` package `mclust` [14].

In this paper we consider a modification of PCA for mixtures with varying concentrations (MVC). The MVC is a nonparametric finite mixture model in which the mixing probabilities (the concentrations of the mixture components) vary from observation to observation. Such models arise naturally in statistical analysis of medical [9] and sociological [12] data. A technique of neuronal activity analysis based on the MVC approach is considered in [13]. See also [2] for adaptive estimation and [1] for adaptive hypotheses testing in MVC models.

In this paper we propose estimators for PC directions and corresponding eigenvectors for each component (subpopulation) of the mixture. Asymptotic normality of these estimators allows one to construct confidence sets for the PC parameters.

The rest of the paper is organized as follows. In Section 2 we give a brief exposition of the classical PC analysis. Section 3 contains general description of the MVC model. In Section 4 we present an estimator for the covariance matrices of the mixture components and derive its asymptotic normality. Section 5 is devoted to the estimators of PC directions and eigenvalues and their asymptotic normality. In Section 6 we apply these results to construction of confidence intervals for the eigenvalues. Section 7 contains results of simulations. The results and further development are discussed in Section 8. One technical result is placed in the Appendix.

2 Classical principal component analysis

Here and below for any univariate sample $\mathbf{x} = (x_1, \dots, x_n)$,

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

denotes the sample mean,

$$S^2(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{\mathbf{x}})^2$$

is the sample variance of $\mathbf{x}$, and $|\mathbf{v}|$ denotes the Euclidean norm of $\mathbf{v}$.

Let $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)$ be a sample of d-dimensional vectors $\mathbf{X}_j = (X_j^1, \dots, X_j^d)^T$, $j = 1, \dots, n$,

$$\mathbf{u}^T\mathbb{X} = (\mathbf{u}^T\mathbf{X}_1, \dots, \mathbf{u}^T\mathbf{X}_n).$$

The first PC direction $\mathbf{v}_1 = \mathbf{v}(1; \mathbb{X})$ of the sample $\mathbb{X}$ is the vector in $\mathbb{R}^d$ of unit length such that

$$S^2(\mathbf{v}_1^T \mathbb{X}) = \max\{S^2(\mathbf{u}^T \mathbb{X}) \ : \mathbf{u} \in \mathbb{R}^d, |\mathbf{u}| = 1\}.$$

Then, for $l = 2, \ldots, d$ the l-th PC direction is the vector $\mathbf{v}_l = \mathbf{v}(l; \mathbb{X})$ of unit length such that

$$S^2(\mathbf{v}_l^T \mathbb{X}) = \max\{S^2(\mathbf{u}^T \mathbb{X}) : \mathbf{u} \in \mathbb{R}^d, |\mathbf{u}| = 1, \mathbf{u}^T \mathbf{v}(1; \mathbb{X}) = 0, \ldots, \mathbf{u}^T \mathbf{v}(l-1; \mathbb{X}) = 0\}.$$

So, the first PC direction is the direction of maximal scattering of a data cloud, the second one is the direction orthogonal to the first one in which the scattering is maximal, and so on.

It is well known that $\mathbf{v}(1, \mathbb{X}), \mathbf{v}(2, \mathbb{X}), \ldots, \mathbf{v}(d, \mathbb{X})$ are the eigenvectors of the sample covariance matrix $\hat{\mathbf{C}}_n = \mathrm{Cov}(\mathbb{X})$ corresponding to its eigenvalues

$$\lambda(1; \mathbb{X}) \geq \lambda(2; \mathbb{X}) \geq \cdots \geq \lambda(d; \mathbb{X}),$$

i.e.

$$\hat{\mathbf{C}}_n \mathbf{v}(l; \mathbb{X}) = \lambda(l; \mathbb{X}) \mathbf{v}(l; \mathbb{X}).$$

Note that

$$\lambda(l; \mathbb{X}) = S^2(\mathbf{v}(l; \mathbb{X})^T \mathbb{X}).$$

If all the eigenvalues are different, then the PC directions are defined unambiguously (up to multiplication by ± 1).

Assume that $\mathbf{X}_j$ are i.i.d. random vectors with a distribution F, i.e. $\mathsf{P}\{\mathbf{X}_j \in A\} = F(A)$ for all Borel sets $A \subseteq \mathbb{R}^d$. Then the PC directions and corresponding eigenvalues of the sample can be interpreted as estimators of the true theoretical PC directions $\mathbf{v}(l; F)$ and eigenvalues $\lambda(l; F)$ which are the eigenvalues and eigenvectors of the covariance of a random vector $\mathbf{X}$ with the distribution F:

$$\mathbf{C} = \mathbf{C}(F) = \mathsf{E}(\mathbf{X} - \mathsf{E}\,\mathbf{X})(\mathbf{X} - \mathsf{E}\,\mathbf{X})^T,$$
$$\mathbf{C}\mathbf{v}(l; F) = \lambda(l; F)\mathbf{v}(l; F), \ l = 1, \ldots, d.$$

These theoretical PC directions possess optimal qualities similar to the sample PC. E.g., the projection of $\mathbf{X}$ on the first PC direction is of maximal variance:

$$\mathrm{Var}(\mathbf{v}(1, F)^T \mathbf{X}) = \max\{\mathrm{Var}(\mathbf{u}^T \mathbf{X}) \ : \ \mathbf{u} \in \mathbb{R}^d, |\mathbf{u}| = 1\}.$$

3 Mixtures with varying concentrations

Now consider a sample of n subjects taken from M different subpopulations (mixture components). For the j-th subject the vector of d observed variables is denoted by $\mathbf{X}_j = (X_j^1, \ldots, X_j^d)^T$. The true number of the component, to which the j-th subject belongs, is denoted by $\kappa_j \in \{1, \ldots, M\}$. This number is not observed, but one knows the probabilities

$$p_j^m = p_{j;n}^m = \mathsf{P}\{\kappa_j = m\}, \ m = 1, \ldots, M.$$

Distribution of $\mathbf{X}_j$ depends on κ_j:

$$F_m(A) = \mathsf{P}\{X_j \in A \mid \kappa_j = m\}.$$

So the unconditional distribution of the observed $\mathbf{X}_j$ is a mixture of M components' distributions

$$\mathsf{P}\{\mathbf{X}_j \in A\} = \sum_{m=1}^{M} p_j^m F_m(A). \tag{1}$$

We assume that $(\mathbf{X}_j, \kappa_j)$ are independent for different $j = 1, \dots, n$. The formula (1) is called the model of mixture with varying concentrations (MVC model). Note that the components' distributions F_m are completely unknown and the concentrations p_j^m are known in this model.

The weighted empirical distribution of the form

$$\hat{F}_{m;n}(A) = \sum_{j=1}^{n} w_j^m \mathbb{1}\{\mathbf{X}_j \in A\} \tag{2}$$

can be used to estimate $F_m(A)$. Here w_j^m are some weights constructed from the concentrations p_j^k, $j = 1, \dots, n$, $k = 1, \dots, M$. These weights are aimed to pick out the m-th component and to suppress the influence of all other components on the estimator.

Investigating the asymptotic behavior of the estimators as the sample size n tends to infinity we will consider different p_j^m and w_j^m for different n. Sometimes it will be denoted by the subscript $_{;n}$: $p_{j;n}^m$, $w_{j;n}^m$, $\mathbf{X}_{j;n}$. If this subscript is dropped it means that we consider here a sample of fixed size n.

Let

$$\mathbf{p}_{;n}^m = (p_{1;n}^m, \dots, p_{n;n}^m)^T, \mathbf{p}_{;n} = (\mathbf{p}_{;n}^1, \dots \mathbf{p}_{;n}^M),$$

i.e. $\mathbf{p}_{;n}$ denotes a matrix of concentrations with M columns and n rows. Each column corresponds to a mixture component, each row corresponds to an observation. Similar notations $\mathbf{w}_{;n}$, $\mathbf{w}_{;n}^m$ will be used for the weights.

Suppose that the vectors $\mathbf{p}_{;n}^m$, $m = 1, \dots, M$, are linearly independent. Then the matrix $\mathbf{\Gamma}_{;n} = (\mathbf{p}_{;n}^T \mathbf{p}_{;n})$ is nonsingular. We will use the weights

$$\mathbf{w}_{;n} = \mathbf{\Gamma}_{;n}^{-1} \mathbf{p}_{;n}. \tag{3}$$

It is shown in [9] that $\hat{F}_{m;n}$ defined by (2) with the weights $\mathbf{w}_{;n}^m$ defined by (3) is a minimax estimator for F_m with respect to the quadratic loss. So the weights (3) are called the minimax weights.

There can be some other choices of weights in (2). E.g., in [10] an adaptive approach is proposed which allows to obtain asymptotically optimal estimators of MVC model parameters. Unfortunately, the adaptive estimators need samples with quite large number observations to outperform the minimax ones in MVC models. Especially large samples are needed for multivariate data analysis which is just the case when PCA is most useful. So, in this paper we will consider estimators with the minimax weights only.

4 Estimation of covariance matrices

In this section we consider estimation of covariance matrices of the mixture components. Assume that $\mathsf{E}[|\mathbf{X}_j|^2 \mid \kappa_j = m] < \infty$ for all $m = 1, \ldots, M$ and let

$$\mu(i; m) = \mathsf{E}[X_j^i \mid \kappa_j = m], \, c(i_1, i_2; m) = \mathsf{E}[X_j^{i_1} X_j^{i_2} - \mu(i_1; m)\mu(i_2; m) \mid \kappa_j = m], \tag{4}$$

$$\mathbf{C}_m = (c(i_1, i_2; m))_{i_1, i_2 = 1}^d.$$

So, $\mathbf{C}_m$ is the covariance matrix of a random vector with the distribution F_m. (The m-th component covariance).

To estimate $\mu(i; m)$ and $c(i_1, i_2; m)$ one can use weighted means with weights designed for the estimation of F_m. Say,

$$\hat{\mu}_{;n}(i; m) = \sum_{j=1}^n w_{j;n}^m X_j^i, \tag{5}$$

$$\hat{c}_{;n}(i_1, i_2; m) = \sum_{j=1}^n w_{j;n}^m X_j^{i_1} X_j^{i_2} - \hat{\mu}_{;n}(i_1; m)\hat{\mu}_{;n}(i_2; m), \tag{6}$$

$$\hat{\mathbf{C}}_{m;n} = (\hat{c}(i_1, i_2; m))_{i_1, i_2 = 1}^d.$$

Theorem 1. Assume that

(i) $\mathsf{E}[|\mathbf{X}_j|^2 \mid \kappa_j = m] < \infty$ for all $m = 1, \ldots, M$;

(ii) There exists nonsingular limit matrix

$$\mathbf{\Gamma}_\infty = \lim_{n \to \infty} \frac{1}{n} \mathbf{\Gamma}_{;n}. \tag{7}$$

Then $\hat{\mathbf{C}}_{m;n} \to \mathbf{C}_m$ in probability.

Proof. This theorem is a simple consequence of the theorem 4.2 in [9]. □

So, under suitable assumptions, $\hat{\mathbf{C}}_{m;n}$ is a consistent estimator for the m-th component covariance matrix. To establish its asymptotic normality we need some additional notations.

Let

$$\begin{aligned} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle_{;n} &= n \sum_{j=1}^n w_{j;n}^{k_1} w_{j;n}^{k_2} p_{j;n}^{m_1} p_{j;n}^{m_2}, \\ \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle &= \lim_{n \to \infty} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle_{;n}, \end{aligned} \tag{8}$$

$$\langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle_{;n} = n \sum_{j=1}^n w_{j;n}^{k_1} w_{j;n}^{k_2} p_{j;n}^{m}, \; \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle = \lim_{n \to \infty} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle_{;n}, \tag{9}$$

assuming that these limits exist. Then

$$\begin{gathered}
\eta_j(i_1, i_2; k) = X_j^{i_1} X_j^{i_2} - X_j^{i_1} \mu(i_2; k) - X_j^{i_2} \mu(i_1; k), \\
M_1(i_1, i_2; k; m) = \mathsf{E}[\eta_j(i_1, i_2; k) \mid \kappa_j = m], \\
M_2(i_1, i_2, i_3, i_4; k_1, k_2; m) = \mathsf{E}[\eta_j(i_1, i_2; k_1)\eta_j(i_3, i_4; k_2) \mid \kappa_j = m], \\
V(i_1, i_2, i_3, i_4; k_1, k_2) = \sum_{m=1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^m \rangle M_2(i_1, i_2, i_3, i_4; k_1, k_2; m) \\
- \sum_{m_1, m_2 = 1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle M_1(i_1, i_2, ; k_1; m_1) M_1(i_3, i_4, ; k_2; m_2).
\end{gathered} \tag{10}$$

Theorem 2. Assume that the following conditions hold.

(i) $\mathsf{E}[|\mathbf{X}_j|^4 \mid \kappa_j = m] < \infty$ for all $m = 1, \dots, M$.

(ii) The matrix $\mathbf{\Gamma}_\infty$ defined by (7) exists and is nonsingular.

(iii) For all $m_1, m_2, k_1, k_2 = 1, \dots, M$ there exist $\langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle$ defined by (8).

Then

$$\sqrt{n}(\hat{\mathbf{C}}_{m;n} - \mathbf{C}_m) \xrightarrow{\mathrm{W}} \mathbf{Z}_m, \text{ for } m = 1, \dots, M,$$

where $\mathbf{Z}_m = (z(i_1, i_2; m))_{i_1, i_2 = 1}^d$, $m = 1, \dots, M$ is a set of matrices with zero mean entries and the covariance structure

$$\mathsf{E}\, z(i_1, i_2; k_1) z(i_3, i_4; k_2) = V(i_1, i_2, i_3, i_4; k_1, k_2),$$

for $j_1, j_2, j_3, j_4 = 1, \dots, d$, $k_1, k_2 = 1, \dots, M$.

Proof. Let

$$\begin{aligned}
z_{;n}(i_1, i_2; k) &= \sqrt{n}(\hat{c}_{;n}(i_1, i_2; k) - c(i_1, i_2; k)), \\
\tilde{z}_{;n}(i_1, i_2; k) &= \sqrt{n} \sum_{j=1}^{n} w_{j;n}^k [\eta_j(i_1, i_2; k) - \mathsf{E}\, \eta_j(i_1, i_2; k)].
\end{aligned} \tag{11}$$

By somewhat tedious but straightforward algebra one obtains

$$\begin{aligned}
&z_{;n}(i_1, i_2; k) - \tilde{z}_{;n}(i_1, i_2; k) \\
&\quad = \sqrt{n} \left(\sum_{j=1}^{n} w_{j;n}^k X_j^{i_1} - \mu(i_1; k) \right) \left(\sum_{j=1}^{n} w_{j;n}^k X_j^{i_2} - \mu(i_2; k) \right).
\end{aligned}$$

Observe that

$$\mathsf{E} \left[\sum_{j=1}^{n} w_{j;n}^k X_j^i \right] = \sum_{j=1}^{n} \sum_{m=1}^{M} p_{j:n}^m w_{j;n}^k \mu(i; m) = \mu(i; k),$$

since $\sum_{j=1}^n p_{j:n}^m w_{j;n}^k$ equals 1 if $m = k$, and 0 otherwise.

So

$$\mathsf{E}\Big[\sum_{j=1}^{n} w_{j;n}^{k} X_j^i - \mu(i;k)\Big]^2 = \sum_{j=1}^{n} (w_{j;n}^{k})^2 \operatorname{Var} X_j^i.$$

Due to the assumption (ii), $\sup_{j,n} \operatorname{Var} X_j^i < \infty$. By lemma 1 in [11], $\sup_{j,n} |w_{j;n}^{k}| = O(n^{-1})$, so, by the Chebyshev inequality,

$$\sum_{j=1}^{n} w_{j;n}^{k} X_j^i - \mu(i;k) = O_P(n^{-1/2})$$

and

$$z_{;n}(i_1, i_2; k) - \tilde{z}_{;n}(i_1, i_2; k) = o_P(1).$$

So it is enough to prove the statement of the Theorem for $\tilde{z}_{;n}(i_1, i_2; k)$ instead of $z_{;n}(i_1, i_2; k)$. Asymptotic normality of the set $(\tilde{z}_{;n}(i_1, i_2; k), i_1, i_2 = 1, \dots, d, k = 1, \dots, M)$ can be proved applying the Central Limit Theorem with the Lindeberg's condition by the same way as in Theorem 4.3. in [9].

Let us calculate the limit covariance. Observe that

$$\mathsf{E}\,\eta_j(i_1, i_2; k) = \sum_{m=1}^{M} p_{j;n}^{m} M_1(i_1, i_2; k; m)$$

and

$$\mathsf{E}\,\eta_j(i_1, i_2; k_1)\eta_j(i_3, i_4; k_2) = \sum_{m=1}^{M} p_{j;n}^{m} M_2(i_1, i_2, i_3, i_4; k_1, k_2; m).$$

So

$$\begin{aligned}
&\mathsf{E}\,\tilde{z}_{;n}(i_1, i_2; k_1)\tilde{z}_{;n}(i_3, i_4; k_2) = \\
&\quad = n \sum_{j=1}^{n} w_{j;n}^{k_1} w_{j;n}^{k_2} \mathsf{E}\,\eta_j(i_1, i_2; k_1)\eta_j(i_3, i_4; k_2) \\
&\quad - n \sum_{j=1}^{n} w_{j;n}^{k_1} w_{j;n}^{k_2} \mathsf{E}\,\eta_j(i_1, i_2; k_1)\,\mathsf{E}\,\eta_j(i_3, i_4; k_2) \\
&\quad = \sum_{m=1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle_{;n} M_2(i_1, i_2, i_3, i_4; k_1, k_2; m) \\
&\quad - \sum_{m_1, m_2 = 1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle_{;n} M_1(i_1, i_2, ; k_1; m_1) M_1(i_3, i_4, ; k_2; m_2).
\end{aligned}$$

Note that, since $\sum_{m=1}^{M} p_{j;n}^{m} = 1$, assumption (iii) implies that

$$\langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle_{;n} = \sum_{k=1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \mathbf{p}^{k} \rangle_{;n} \to \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \rangle = \sum_{k=1}^{M} \langle \mathbf{w}^{k_1} \mathbf{w}^{k_2} \mathbf{p}^{m} \mathbf{p}^{k} \rangle$$

as $n \to \infty$. So

$$\mathsf{E}\,\tilde{z}_{;n}(i_1, i_2; k_1)\tilde{z}_{;n}(i_3, i_4; k_2) \to V(i_1, i_2, i_3, i_4; k_1, k_2).$$

The Theorem is proved. □

To apply this theorem for the construction of confidence interval or hypotheses testing one needs an estimator for the asymptotic covariances $V(i_1, i_2, i_3, i_4; k_1, k_2)$. To obtain it, let us consider

$$\hat{\eta}_{j;n}(i_1, i_2; k) = X_j^{i_1} X_j^{i_2} - X_j^{i_1}\hat{\mu}_{;n}(i_2; k) - X_j^{i_2}\hat{\mu}_{;n}(i_1; k).$$

Observe that, under the assumptions of Theorem 2,

$$\widehat{M}_{1;n}(i_1, i_2; k; m) = \sum_{j=1}^{n} w_{j;n}^{m}\hat{\eta}_{j;n}(i_1, i_2; k)$$

and

$$\widehat{M}_{2;n}(i_1, i_2, i_3, i_4; k_1, k_2; m) = \sum_{j=1}^{n} w_{j;n}^{m}\hat{\eta}_{j;n}(i_1, i_2; k_1)\hat{\eta}_{j;n}(i_3, i_4; k_2)$$

are consistent estimators to $M_1(i_1, i_2; k; m)$ and $M_2(i_1, i_2, i_3, i_4; k_1, k_2; m)$ respectively. So

$$\begin{aligned}\widehat{V}_{;n}(i_1, i_2, i_3, i_4; k_1, k_2) &= \sum_{m=1}^{M}\langle \mathbf{w}^{k_1}\mathbf{w}^{k_2}\mathbf{p}^{m}\rangle_{;n}\widehat{M}_{2;n}(i_1, i_2, i_3, i_4; k_1, k_2; m) \\ &- \sum_{m_1, m_2=1}^{M}\langle \mathbf{w}^{k_1}\mathbf{w}^{k_2}\mathbf{p}^{m_1}\mathbf{p}^{m_2}\rangle_{;n}\widehat{M}_{1;n}(i_1, i_2, ; k_1; m_1)\widehat{M}_{1;n}(i_3, i_4, ; k_2; m_2).\end{aligned} \tag{12}$$

is a consistent estimator to $V(i_1, i_2, i_3, i_4; k_1, k_2)$.

5 Principal components for mixtures

We define the principal components directions of the k-th mixture component as the eigenvectors of $\mathbf{C}_k$. Let $\lambda(1; k) > \lambda(2; k) > \cdots \lambda(d; k)$ be the eigenvalues of $\mathbf{C}_k$ and $\mathbf{v}(l; k) = (v^1(l; k), \ldots, v^d(l; k))^T$ be the corresponding eigenvectors:

$$\mathbf{C}_k\mathbf{v}(l; k) = \lambda(l; k)\mathbf{v}(l; k). \tag{13}$$

In what follows we assume that all the eigenvalues of $\mathbf{C}_k$ are simple (i.e. there are d different eigenvalues) and $|\mathbf{v}(l; k)| = 1$. Then these vectors are defined unambiguously up to the sign multiplier ± 1: if (13) holds for $\mathbf{v}(l; k)$ then $-\mathbf{v}(l; k)$ also satisfy it.

To avoid the ambiguity, we adopt the following rule for choosing the sign of an eigenvector. Consider $v = \max_{i=1,\ldots,d}|v^i(l; k)|$ and $i_0 = \min\{i : |v^i(l; k)| = v\}$. We choose as the PC direction the version of $\mathbf{v}(l; k)$ for which $v^{i_0}(l; k) > 0$.

Natural estimators for $\lambda(l;k)$ and $\mathbf{v}(l;k)$ are the eigenvalues and eigenvectors of $\hat{\mathbf{C}}_{k;n}$. Let $\hat{\lambda}_{;n}(l;k)$ denote the l-th (in the decreasing order) eigenvalue of $\hat{\mathbf{C}}_{k;n}$. To choose the sign of the corresponding estimated eigenvector $\hat{\mathbf{v}}_{;n}(l;k) = (\hat{v}^1_{;n}(l;k), \ldots, \hat{v}^d_{;n}(l;k))^T$ we need somewhat more complicated algorithm than in the case of $\mathbf{v}(l;k)$.

Let $\varepsilon_n > 0$ be some sequence such that $\varepsilon_n \to 0$ as $n \to \infty$. Consider $\hat{v}_{;n} = \max_{i=1,\ldots,d} |\hat{v}^i_{;n}(l;k)|$ and $\hat{i}_0 = \min\{i \ : |\hat{v}^i(l;k)| \geq \hat{v}_{;n} - \varepsilon_n\}$. Then we choose the sign of $\hat{\mathbf{v}}_{;n}(l;k)$ so that $\hat{v}^{\hat{i}_0}_{;n} > 0$.

Let $\mathbb{E}$ be the unit $d \times d$-matrix and $\mathbf{A}^+$ denotes the Moore–Penrose inverse of a matrix $\mathbf{A}$.

Theorem 3. Assume the following.

(i) $\mathsf{E}[|\mathbf{X}_j|^4 \mid \kappa_j = m] < \infty$ for all $m = 1, \ldots, M$.

(ii) The matrix $\mathbf{\Gamma}_\infty$ defined by (7) exists and is nonsingular.

(iii) For all $m_1, m_2, k_1, k_2 = 1, \ldots, M$ there exist $\langle \mathbf{w}^{k_1}\mathbf{w}^{k_2}\mathbf{p}^{m_1}\mathbf{p}^{m_2}\rangle$ defined by (8).

(iv) All the eigenvalues of $\mathbf{C}_k$ are simple.

(v) $\varepsilon_n \to 0$, $\sqrt{n}\varepsilon_n \to \infty$ as $n \to \infty$.

Then

$$\sqrt{n}(\hat{\lambda}_{;n}(l;k)) - \lambda(l;k)) \xrightarrow{\mathrm{W}} \mathbf{v}^T(l;k)\mathbf{Z}_k\mathbf{v}(l;k), \tag{14}$$

$$\sqrt{n}(\hat{\mathbf{v}}_{;n}(l;k)) - \mathbf{v}(l;k)) \xrightarrow{\mathrm{W}} (\mathbf{C}_k - \lambda(l;k)\mathbb{E})^+\mathbf{Z}_k\mathbf{v}(l;k), \tag{15}$$

as $n \to \infty$, for all $l = 1, \ldots, d$, where $\mathbf{Z}_k$ is defined in Theorem 2.

Remark. The weak convergence in (14)–(15) is simultaneous. I.e. the common distribution of the set of LHS for $l = 1, \ldots, d$ in (14)–(15) converges to the common distribution of RHS. Moreover, if assumption (iv) holds for some set of k, then the convergence is simultaneous for these k.

Proof. Consider the l-th eigenvalue $\lambda_l(\mathbf{C})$ of a symmetric matrix $\mathbf{C} \in \mathbb{R}^{d\times d}$ as a function of $\mathbf{C}$. Since all the eigenvalues of $\mathbf{C}_k$ are simple, there exists a neighborhood N_k of $\mathbf{C}_k$ at which $\lambda_l(\mathbf{C})$ is continuous. It is well known that this is a continuously differentiable function of $\mathbf{C}$. There exist also two versions $\pm\mathbf{v}_l(\mathbf{C})$ of the l-th eigenvector of $\mathbf{C}$ each of which is continuously differentiable in N_k. We choose the version $\mathbf{v}_l(\mathbf{C})$ for which $\mathbf{v}_l(\mathbf{C}_k) = \mathbf{v}(l,k)$.

Consider a continuously differentiable parametric family $\mathbf{C}_t$, $t \in (a,b) \subset \mathbb{R}$, of $d \times d$ symmetric matrices with simple eigenvalues. Differentiating the equations

$$\mathbf{C}_t\mathbf{v}_l(\mathbf{C}_t) = \lambda_l(\mathbf{C}_t)\mathbf{v}_l(\mathbf{C}_t)$$

and

$$(\mathbf{v}_l(\mathbf{C}_t))^T\mathbf{v}_l(\mathbf{C}_t) = 1,$$

one obtains

$$\frac{d}{dt}\lambda_l(\mathbf{C}_t) = (\mathbf{v}_l(\mathbf{C}_t))^T \left(\frac{d}{dt}\mathbf{C}_t\right) \mathbf{v}_l(\mathbf{C}_t), \tag{16}$$

$$\frac{d}{dt}\mathbf{v}_l(\mathbf{C}_t) = (\mathbf{C}_t - \lambda_l(\mathbf{C}_t)\mathbb{E})^+ \left(\frac{d}{dt}\mathbf{C}_t\right) \mathbf{v}_l(\mathbf{C}_t) \tag{17}$$

(Theorem 8.9 in [8]). By the Taylor expansion

$$\hat{\lambda}_{;n}(l;k)) - \lambda(l;k) = \lambda_l(\hat{\mathbf{C}}_{k;n}) - \lambda_l(\mathbf{C}_k)$$

$$= \sum_{i,j=1}^{d} \left(\frac{\partial}{\partial c(i,j;k)}\lambda_l(\mathbf{C}_k)\right) (\hat{c}_{;n}(i,j;k) - c(i,j;k)) + o(|\hat{\mathbf{C}}_{k;n} - \mathbf{C}_k|).$$

Applying (16) with $t = c(i, j; k)$ we obtain

$$\sqrt{n}(\hat{\lambda}_{;n}(l;k)) - \lambda(l;k)) = (\mathbf{v}(l;k))^T \sqrt{n}(\hat{\mathbf{C}}_{k;n} - \mathbf{C}_k)\mathbf{v}(l;k) + o(|\hat{\mathbf{C}}_{k;n} - \mathbf{C}_k|).$$

By Theorem 2 this implies (14).

Proceeding by the same way with (17) in view we obtain

$$\sqrt{n}(\mathbf{v}_l(\hat{\mathbf{C}}_{k;n}) - \mathbf{v}_l(\mathbf{C}_k)) \xrightarrow{\mathrm{W}} (\mathbf{C}_k - \lambda(l;k)\mathbb{E})^+ \mathbf{Z}_k \mathbf{v}(l;k).$$

It is rather similar to (15), but recall that $\hat{\mathbf{v}}_{;n}(l;k)) = \pm\mathbf{v}_l(\hat{\mathbf{C}}_{k;n})$. Observe that $\mathbf{v}_l(\hat{\mathbf{C}}_{k;n}) - \mathbf{v}_l(\mathbf{C}_k) = O_P(1/\sqrt{n})$, while $1/\sqrt{n} = o(\varepsilon_n)$. So our sign choosing rule for $\hat{\mathbf{v}}_{;n}(l;k)$ will choose the right sign with probability tending to 1 as $n \to \infty$:

$$\mathsf{P}\{\hat{\mathbf{v}}_{;n}(l;k) = \mathbf{v}_l(\hat{\mathbf{C}}_{k;n})\} \to 1 \text{ as } n \to \infty.$$

Theorem is proved. □

6 Confidence intervals for eigenvalues

Theorem 3 can be applied to testing hypotheses on PC directions of different mixture components. It also allows one to construct confidence sets for PC directions and eigenvalues. As an example we consider construction of a confidence interval for one eigenvalue $\lambda(l;k)$ of the k-th mixture component covariance matrix $\mathbf{C}_k$.

By Theorem 3

$$\sqrt{n}(\hat{\lambda}_{;n}(l;k) - \lambda(l;k)) \xrightarrow{\mathrm{W}} N(0, S^2(l,k)),$$

where

$$S^2(l,k) = \mathrm{Var}\,\mathbf{v}^T(l;k)\mathbf{Z}_k\mathbf{v}(l;k)$$

$$= \sum_{i_1,i_2,i_3,i_4=1}^{d} v^{i_1}(l;k)v^{i_2}(l;k)v^{i_3}(l;k)v^{i_4}(l;k)V(i_1,i_2,i_3,i_4;k,k). \tag{18}$$

An estimator $\hat{S}^2_{;n}(l;k)$ for $S^2(l;k)$ can be obtained by replacing $v^i(l;k)$ and $V(i_1,i_2,i_3,i_4;k,k)$ by their consistent estimators $\hat{v}^i_{;n}(l;k)$ and $\hat{V}_{;n}(i_1,i_2,i_3,i_4;k,k)$. It is obvious that $\hat{S}^2_{;n}(l;k)$ is consistent under the assumptions of Theorem 3.

So, if $S^2(l,k) > 0$, then

$$\frac{\sqrt{n}}{\hat{S}_{;n}(l;k)}(\hat{\lambda}_{;n}(l;k) - \lambda(l;k)) \xrightarrow{\mathrm{W}} N(0,1).$$

Let $x_{\alpha/2}$ be the standard normal quantile of level $1-\alpha/2$,

$$\lambda^{\pm}_{;n}(l;k) = \hat{\lambda}_{;n}(l;k) \pm x_{\alpha/2}\frac{\hat{S}_{;n}(l;k)}{\sqrt{n}}.$$

Then

$$\lim_{n\to\infty} \mathsf{P}\{\lambda(l;k) \in [\lambda^{-}_{;n}(l;k), \lambda^{+}_{;n}(l;k)]\} = 1-\alpha.$$

I.e. $[\lambda^{-}_{;n}(l;k), \lambda^{+}_{;n}(l;k)]$ is an asymptotic confidence interval for $\lambda(l;k)$ with the significance level α if the assumptions of Theorem 3 hold and $S^2 > 0$. The last assumption is not too restrictive. In the Appendix we present a simple condition under which it holds.

7 Results of simulations

To evaluate the finite-sample behavior of the proposed technique, we performed a small simulation study. Confidence intervals for the largest eigenvalue $\lambda(1;k)$ with the nominal significance level $\alpha = 0.05$ were calculated on data simulated from the tree-component MVC model. In each experiment there were $B = 1000$ simulations for each sample size $n = 250, 500, \ldots, 10000$.

For each mixture component we present the coverage frequency of the intervals, i.e. the number of confidence intervals which covered the true $\lambda(1;k)$ divided by B.

In all the experiments the concentrations $\mathbf{p}_{j;n} = (p^1_{j;n}, p^2_{j;n}, p^3_{j;n})$ were generated as independent vectors uniformly distributed on the simplex $\{\mathbf{p} : p^m \geq 0, m = 1,\ldots,3, p^1+p^2+p^3 = 1\}$.

The observations $\mathbf{X}_j = (X^1_j, X^2_j, X^3_j)$ were three-dimensional.

In the **first experiment** the distribution of $\mathbf{X}_j$ was Gaussian in each mixture component $F_m \sim N(\boldsymbol{\mu}_m, \mathbf{C}_m)$, where

$$\boldsymbol{\mu}_1 = \begin{pmatrix}1\\0\\2\end{pmatrix},\ \boldsymbol{\mu}_2 = \begin{pmatrix}0\\0\\0\end{pmatrix},\ \boldsymbol{\mu}_3 = \begin{pmatrix}1\\2\\3\end{pmatrix} \tag{19}$$

and

$$\mathbf{C}_1 = \begin{pmatrix}1 & -0.5 & 0.1\\ -0.5 & 2 & 0.4\\ 0.1 & 0.4 & 3\end{pmatrix},\ \mathbf{C}_2 = \begin{pmatrix}2 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0.5\end{pmatrix},\ \mathbf{C}_3 = \begin{pmatrix}5 & 1 & 1\\ 1 & 2 & 1\\ 1 & 1 & 0.5\end{pmatrix}. \tag{20}$$

Table 1. Coverage probabilities of the first experiment

	Components		
n	first	second	third
250	0.973	0.872	0.956
500	0.969	0.925	0.952
1000	0.962	0.968	0.953
2500	0.952	0.966	0.951
5000	0.948	0.941	0.956
10000	0.955	0.960	0.953

Table 2. Coverage probabilities of the second experiment

	Components		
n	first	second	third
250	0.971	0.839	0.959
500	0.961	0.879	0.964
1000	0.964	0.919	0.955
2500	0.954	0.925	0.945
5000	0.943	0.932	0.945
10000	0.950	0.921	0.947

The simulation results are presented on Table 1. It seems that the coverage probabilities of obtained confidence intervals are satisfactory for practical purposes if the sample size n is greater then 1000.

In the **second experiment** we used the same parameters for the model as in the first one, except the covariance of the second component. Here we put $\mathbf{C}_2 = \mathbb{E}$ (unit matrix). Surely, this model does not satisfy the assumptions of Theorem 3, since the first eigenvalue of $\mathbf{C}_2$ is not simple. The results are presented on Table 2. It seems that the confidence interval for the second component's largest eigenvalue is unsatisfactory even for $n = 10000$. The intervals for the first and third components preform satisfactory for n larger then 1000.

8 Discussion

We proposed a technique for estimation of PC directions and eigenvalues by observations from MVC. Asymptotic normality of the estimators is proved. This opens possibilities for constructing confidence sets and testing hypotheses on PC structure of different mixture components. Results of simulations confirm applicability of the asymptotic results for samples of moderate size.

Now let us discuss some challenges which were not answered in this study.

1. To apply Theorem 3 for statistical analysis of real life data, one needs to be sure that the assumption (iv) of this theorem holds. Is it possible to verify the hypotheses A: *all the eigenvalues of* $\mathbf{C}_k$ *are simple* by some statistical test? Note, that in this test A must be an alternative to the null H_0: *there exists an eigenvalue of* $\mathbf{C}_k$ *of degree higher then 1*. Of course, such test cannot be based on the confidence sets derived by Theorem 3, since it does not hold under H_0. A possible alternative is to use a bootstrap technique for such testing.

2. It is sometimes useful in cluster analysis applications to consider FMMs with growing number of components (clusters) as the sample size n tends to infinity [5]. Similar approach is also beneficial in signal processing [3]. In this case one expects nonparametric convergence rates in theorems similar to Theorem 3. Another generalizations of Theorem 3 are possible if the dimension of the observations space $d \to \infty$ as $n \to \infty$.

3. There are many alternatives to PCA as a dimension reduction technique, e.g., Projection Pursuit (PP) or Independent Components Analysis [6]. Some of them, such as the PP based on the maximization of kurtosis can be modified for application to MVC data similarly to the PCA modification considered in this study. It would be interesting to analyze efficiency of these modifications both theoretically and in real life data analysis.

We hope that further study will clarify answers on these questions.

A Appendix

Here we will obtain conditions under which $S^2(l;k)$ defined by (18) is strictly positive.

Let vect be a function which stacks its arguments into a long vector:

$$\mathrm{vect}(z_{ij},\ i,j=1,\dots,d)=\mathbf{z}=(z^1,\dots,z^{d^2}) \text{ where } z^{i+d(j-1)}=z_{ij}.$$

Let us fix $k \in 1,\dots,M$ and define $\boldsymbol{\eta}_j(k)=\mathrm{vect}(\eta_j(i_1,i_2;k),\ i_1,i_2=1,\dots d)$.

Theorem 4. Let the assumptions of Theorem 3 hold. If for some $m \in 1,\dots,M$

$$\det \mathrm{Cov}[\boldsymbol{\eta}_j \mid \kappa_j = m] \neq 0,$$

then $S^2(l;k) > 0$ for all $l = 1,\dots,d$.

Proof. Let

$$\mathbf{z}=\mathrm{vect}(z(i_1,i_2;k),\ i_1,i_2=1,\dots,d),\ \mathbf{v}=\mathrm{vect}(v^{i_1}(l;k)v^{i_2}(l;k),\ i_1,i_2=1,\dots,d).$$

By (18)

$$S^2(l;k)=\mathbf{v}^T\,\mathrm{Cov}[\mathbf{z}]\mathbf{v}.$$

It is obvious that $|\mathbf{v}| \neq 0$. So, to obtain $S^2(l;k) > 0$ we need only to show that $\mathrm{Cov}[\mathbf{z}] > 0$.

Let $\mathbf{e}_m = \mathsf{E}[\boldsymbol{\eta}_j \mid \kappa_j = m]$, then

$$\mathrm{Cov}[\mathbf{z}] = \mathbf{L}_1 + \mathbf{L}_2,$$

where

$$\mathbf{L}_1 = \sum_{m=1}^{M} \langle (\mathbf{w}^k)^2 \mathbf{p}^m \rangle \left(\mathsf{E}[\boldsymbol{\eta}_j \boldsymbol{\eta}_j^T \mid \kappa_j = m] - \mathbf{e}_m \mathbf{e}_m^T \right),$$

$$\mathbf{L}_2 = \sum_{m=1}^{M} \langle (\mathbf{w}^k)^2 \mathbf{p}^m \rangle \mathbf{e}_m \mathbf{e}_m^T - \sum_{m_1,m_2=1}^{M} \langle (\mathbf{w}^k)^2 \mathbf{p}^{m_1} \mathbf{p}^{m_2} \rangle \mathbf{e}_{m_1} \mathbf{e}_{m_2}^T.$$

Observe that $\mathbf{L}_2 = \lim_{n\to\infty} \sum_{j=1}^{n} (w_{j;n}^k)^2 \mathbf{L}_{j,2}$, where

$$\mathbf{L}_{j,2} = \mathrm{Cov}(\boldsymbol{\zeta}_j),$$

$\boldsymbol{\zeta}_j$ is a random vector which attains values $\mathbf{e}_m$ with probabilities $p_{j;n}^m$. So $\mathbf{L}_{j;2} \geq 0$ and $\mathbf{L}_2 \geq 0$.

Then

$$\mathbf{L}_1 = \sum_{j=1}^{n} \langle (\mathbf{w}^k)^2 \mathbf{p}^m \rangle \, \mathrm{Cov}[\boldsymbol{\eta}_j \mid \kappa_j = m] \leq \langle (\mathbf{w}^k)^2 \mathbf{p}^{m_0} \rangle \, \mathrm{Cov}[\boldsymbol{\eta}_j \mid \kappa_j = m_0] > 0,$$

due to the assumpton of the Theorem and the fact that $\langle (\mathbf{w}^k)^2 \mathbf{p}^{m_0} \rangle > 0$ (see [12]). Summarizing we obtain

$$\mathrm{Cov}[\mathbf{z}] > 0$$

which implies the statement of the Theorem. □

Acknowledgement

We are thankful to the unknown referees for their attention to our work, fruitful comments and suggestions.

References

[1] Autin, F., Pouet, C.: Adaptive test on components of densities mixture. Math. Methods Stat. **21**(2), 93–108 (2012). MR2974011. https://doi.org/10.3103/S1066530712020020

[2] Doronin, O.: Adaptive estimation for a semiparametric model of mixture. Theory Probab. Math. Stat. **91**, 29–41 (2015). MR3364121. https://doi.org/10.1090/tpms/964

[3] Gayraud, G., Ingster, Y.I.: Detection of sparse additive functions. Electron. J. Stat. **6**, 1409–1448 (2012). MR2988453. https://doi.org/10.1214/12-EJS715

[4] Härdle, W., Simar, L.: Applied Multivariate Statistical Analysis. Springer, Berlin Heidelberg (2007). MR2367300

[5] Ho, N., Nguyen, X.: Convergence rates of parameter estimation for some weakly identifiable finite mixtures. Ann. Stat. **44**, 2726–2755 (2016). MR3576559. https://doi.org/10.1214/16-AOS1444

[6] Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. Wiley (2001)

[7] Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2010). MR0841268. https://doi.org/10.1007/978-1-4757-1904-8

[8] Magnus, J.R., Neudecker, H.: Matrix Differential Calculus with Applications in Statistics and Econometrics. Wiley, New York (2019). MR1698873

[9] Maiboroda, R., Sugakova, O.: Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. J. Nonparametr. Stat. **24**(1), 201–215 (2012). MR2885834. https://doi.org/10.1080/10485252.2011.630076

[10] Maiboroda, R., Sugakova, O., Doronin, A.: Generalized estimating equations for mixtures with varying concentrations. Can. J. Stat. **41**(2), 217–236 (2013). MR3061876. https://doi.org/10.1002/cjs.11170

[11] Maiboroda, R., Sugakova, O.: Jackknife covariance matrix estimation for observations from mixture. Mod. Stoch. Theory Appl. **6**(4), 495–513 (2019). MR4047396. https://doi.org/10.15559/19-vmsta145

[12] Miroshnichenko, V., Maiboroda, R.: Asymptotic normality of modified LS estimator for mixture of nonlinear regressions. Mod. Stoch. Theory Appl. **7**(4), 435–448 (2020). MR4195645

[13] Pidnebesna, A., Fajnerová, I., Horáček, J., Hlinka, J.: Estimating Sparse Neuronal Signal from Hemodynamic Response: The Mixture Components Inference Approach. https://www.biorxiv.org/content/10.1101/2019.12.19.876508v1. Accessed 8 August 2021.

[14] Scrucca, L., Fop, M., Murphy, T.B., Raftery, A.E.: mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. R J. **8**(1), 289–317 (2016). https://doi.org/10.32614/RJ-2016-021

[15] Van Huffel, S., Vandewalle, J.: The Total Least Squares Problem – Computational Aspects and Analysis. Society for Industrial and Applied Mathematics, Philadelphia (1991). MR1118607. https://doi.org/10.1137/1.9781611971002