DOI: 10.15559/15-VMSTA41

Linear regression by observations from mixture with varying concentrations

Daryna Liubashenko, Rostyslav Maiboroda[*]

Kyiv National Taras Shevchenko University, Kyiv, Ukraine

dlyubashenko@gmail.com (D. Liubashenko), mre@univ.kiev.ua (R. Maiboroda)

Received: 20 October 2015, Revised: 22 November 2015, Accepted: 26 November 2015,
Published online: 4 December 2015

Abstract We consider a finite mixture model with varying mixing probabilities. Linear regression models are assumed for observed variables with coefficients depending on the mixture component the observed subject belongs to. A modification of the least-squares estimator is proposed for estimation of the regression coefficients. Consistency and asymptotic normality of the estimates is demonstrated.

Keywords Finite mixture model, linear regression, mixture with varying concentrations, nonparametric estimation, asymptotic normality, consistency

2010 MSC 62J05, 62G20

1 Introduction

In this paper, we discuss a structural linear regression technique in the context of model of mixture with varying concentrations (MVC). MVC means that the observed subjects belong to M different subpopulations (mixture components). The true numbers of components to which the subjects O_j, $j = 1, \dots, N$, belong, say, $\kappa_j = \kappa(O_j)$, are unknown, but we know the probabilities

$$p^k_{j;N} \overset{\text{def}}{=} \mathsf{P}\{\kappa_j = k\}, \quad k = 1, \dots, M$$

[*]Corresponding author.

(mixing probabilities or concentrations of the mixture components). MVC models arise naturally in the description of medical, biologic, and sociologic data [1, 8, 9, 12]. They can be considered as a generalization of finite mixture models (FMM). Classical theory of FMMs can be found in monographs [10, 13].

Let

$$\xi(O) = \big(Y(O), X^1(O), \dots, X^d(O)\big)^T$$

be a vector of observed features (random variables) of a subject O. We consider the following linear regression model for these variables:

$$Y(O) = \sum_{i=1}^{d} b_i^{(\kappa(O))} X^i(O) + \varepsilon(O), \tag{1}$$

where $\mathbf{b}^{(m)} = (b_1^{(m)}, \dots, b_d^{(m)})^T$ are nonrandom regression coefficients for the mth component, $\varepsilon(O)$ is an error term, which is assumed to be zero mean and conditionally independent of the regressors vector $\mathbf{X}(O) = (X^1(O), \dots, X^d(O))^T$ given $\kappa(O)$.

Note. We consider a subject O as taken at random from an infinite population, so it is random in this sense. The vector of observed variables $\xi(O)$ can be considered as a random vector even for a fixed O.

Our aim is to estimate the vectors of regression coefficients $\mathbf{b}^{(k)}$, $k = 1, \dots, M$, by the observations $\Xi_N = (\xi_1, \dots, \xi_N)$, where $\xi_j = \xi(O_j)$. We assume that (κ_j, ξ_j) are independent for different j.

A statistical model similar to MVC with (1) is considered in [5], where a parametric model for the conditional distributions of $\varepsilon(O)$ given $\kappa(O)$ is assumed. For this case, maximum likelihood estimation is proposed in [5], and a version of EM-algorithm is developed for numerical computation of the estimates.

In this paper, we adopt a nonparametric approach assuming no parametric models for $\varepsilon(O)$ and $\mathbf{X}(O)$ distributions. Nonparametric and semiparametric technique for MVC was developed in [6, 7, 4]. We use the weighted empirical moment technique to derive estimates for the regression coefficients and then obtain conditions of consistency and asymptotic normality of the estimates. These results are based on general ideas of least squares [11] and moment estimates [3].

The rest of the paper is organized as follows. In Section 2, we recall some results on nonparametric estimation of functional moments in general MVC. The estimates are introduced, and conditions of their consistency and asymptotic normality are presented in Section 3. Section 4 contains proofs of the statements of Section 3. Results of computer simulations are presented in Section 5.

2 Nonparametric estimation for MVC

Let us start with some notation and definitions. We denote by F_m the distribution of $\xi(O)$ for O belonging to the mth component of the mixture, that is,

$$F_m(A) \overset{\text{def}}{=} \mathsf{P}\big\{\xi(O) \in A \mid \kappa(O) = m\big\}$$

for all measurable sets A. Then by the definition of MVC

$$\mathsf{P}\{\xi_j \in A\} = \sum_{k=1}^{M} p^k_{j;N} F_k(A). \tag{2}$$

In the asymptotic statements, we will consider the data $\Xi_n = (\xi_1, \dots, \xi_N)$ as an element of (imaginary) series of data $\Xi_1, \Xi_2, \dots, \Xi_N, \dots$ in which no link between observations for different N is assumed. So, in formal notation, it should be more correct to write $\xi_{j;N}$ instead of ξ_j, but we will drop the subscript N when it is insignificant.

We consider an array of all concentrations for all data sizes

$$\mathbf{p} \overset{\text{def}}{=} \big(p^k_{j;N}, k = 1, \dots, M;\ j = 1, \dots, N;\ N = 1, 2, \dots\big).$$

Its subarrays

$$\mathbf{p}^k_N \overset{\text{def}}{=} \big(p^k_{1;N}, \dots, p^k_{N;N}\big)^T \quad \text{and} \quad \mathbf{p}_{j;N} = \big(p^1_{j;N}, \dots, p^M_{j;N}\big)^T$$

are considered as vector columns, and

$$\mathbf{p}_N = \big(p^k_{j;N},\ j = 1, \dots, N;\ k = 1, \dots, M\big)$$

as an $N \times M$-matrix. We will also consider a weight array $\mathbf{a}$ of the same structure as $\mathbf{p}$ with similar notation for its subarrays.

By the angle brackets with subscript n we denote the averaging by $j = 1, \dots, n$:

$$\langle \mathbf{a}^k \rangle_N \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} a^k_{j;N}.$$

Multiplication, summation, and other operations in the angle brackets are made elementwise:

$$\langle \mathbf{a}^k \mathbf{p}^m \rangle_N = \frac{1}{N} \sum_{j=1}^{N} a^k_{j;N} p^m_{j;N}; \qquad \big\langle (\mathbf{a}^k)^2 \big\rangle_N = \frac{1}{N} \sum_{j=1}^{N} \big(a^k_{j;N}\big)^2.$$

We define $\langle \mathbf{p}^m \rangle \overset{\text{def}}{=} \lim_{N\to\infty} \langle \mathbf{p}^m \rangle_N$ if this limit exists. Let $\boldsymbol{\Gamma}_N = (\langle \mathbf{p}^l \mathbf{p}^m \rangle_N)_{l,m=1}^{M} = \frac{1}{N} \mathbf{p}_N^T \mathbf{p}_N$ be an $M \times M$ matrix, and $\gamma_{lm;N}$ be its (l, m)th minor. The matrix $\boldsymbol{\Gamma}_N$ can be considered as the Gramian matrix of vectors $(\mathbf{p}^1, \dots, \mathbf{p}^M)$ in the inner product $\langle \mathbf{p}^i \mathbf{p}^k \rangle_N$, so that it is nonsingular if these vectors are linearly independent.

In what follows, $\xrightarrow{\text{P}}$ means convergence in probability, and $\xrightarrow{\text{W}}$ means weak convergence.

Assume now that model (2) holds for the data Ξ_N. Then the distribution F_m of the mth component can be estimated by the weighted empirical measure

$$\hat{F}_{m;N}(A) \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} a^m_{j;N} \mathbb{1}\{\xi_j \in A\}, \tag{3}$$

where

$$a^m_{j;N} = \frac{1}{\det \boldsymbol{\Gamma}_N} \sum_{k=1}^{M} (-1)^{k+m} \gamma_{mk;N} p^k_{j;N}. \tag{4}$$

It is shown in [8] that if $\boldsymbol{\Gamma}_N$ is nonsingular, then $\hat{F}_{m;N}$ is the minimax unbiased estimate for F_m. The consistency of $\hat{F}_{m;N}$ is demonstrated in [6] (see also [8]).

Consider now functional moment estimation based on weighted empirical moments. Let $g : \mathbb{R}^{d+1} \to \mathbb{R}^k$ be a measurable function. Then to estimate

$$\bar{g}^{(m)} = \mathsf{E}\big[g\big(\xi(O)\big) \mid \kappa(O) = m\big] = \int g(x) F_m(dx),$$

we can use

$$\hat{g}^{(m)}_{;N} = \int g(x) \hat{F}_{m;N}(dx) = \frac{1}{N} \sum_{j=1}^{N} a^m_{j;N} g(\xi_j).$$

Lemma 1 (Consistency)**.** *Assume that*

1. $\mathsf{E}[|g(\xi(O))| \mid \kappa(O) = k] < \infty$ *for all* $k = 1, \ldots, M$.
2. *There exists* $C > 0$ *such that* $\det \boldsymbol{\Gamma}_N > C$ *for all* N *large enough.*

Then $\hat{g}^{(m)}_{;N} \xrightarrow{\mathrm{P}} \bar{g}^{(m)}$ as $N \to \infty$.

This lemma is a simple corollary of Theorem 4.2 in [8]. (See also Theorem 3.1.1 in [7]).

Lemma 2 (Asymptotic normality)**.** *Assume that*

1. $\mathsf{E}[|g(\xi(O))|^2 \mid \kappa(O) = k] < \infty$ *for all* $k = 1, \ldots, M$.
2. *There exists* $C > 0$ *such that* $\det \boldsymbol{\Gamma}_N > C$ *for all* N *large enough.*
3. *There exists the limit*

$$\boldsymbol{\Sigma} = \lim_{N \to \infty} N \operatorname{Cov}\big(\hat{g}^{(m)}_{;N}\big).$$

Then

$$\sqrt{N}\big(\hat{g}^{(m)}_{;N} - \bar{g}^{(m)}\big) \xrightarrow{\mathrm{W}} N(0, \boldsymbol{\Sigma}).$$

For univariate $\hat{g}^{(m)}_{;N}$, the statement of the lemma is contained in Theorem 4.2 from [8] (or Theorem 3.1.2 in [7]). The multivariate case can be obtained from the univariate one applying the Cramér–Wold device (see [2], p. 382).

3 Estimate for $\mathbf{b}_m$ and its asymptotics

In view of Lemma 1, we expect that, under suitable assumptions,

$$J_{m;N}(\mathbf{b}) \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} \mathbf{a}^m_{j;N} \Bigg(Y_{j;N} - \sum_{i=1}^{d} b_i X^i_{j:N} \Bigg)^2$$

$$= \int \Bigg(y - \sum_{i=1}^{d} b_i x^i \Bigg)^2 \hat{F}_{m;N}\big(dy, dx^1, \ldots, dx^d\big)$$

converges to

$$\int\Big(y-\sum_{i=1}^{d} b_i x^i\Big)^2 F_m\big(dy, dx^1, \dots, dx^d\big)$$
$$= \mathsf{E}\Bigg[\Big(Y(O)-\sum_{i=1}^{d} b_i X^i(O)\Big)^2 \;\Bigg|\; \kappa(O)=m\Bigg] \overset{\text{def}}{=} J_{m;\infty}(\mathbf{b})$$

as $N \to \infty$.

Since $J_{m;\infty}(\mathbf{b})$ attains its minimal value at $\mathbf{b}^{(m)}$, it is natural to suggest the argmin of $J_{m;N}(\mathbf{b})$ as an estimate for $\mathbf{b}^{(m)}$. If the weights $\mathbf{a}^m$ were positive, then this argmin would be

$$\hat{\mathbf{b}}^{(m)}_{;N} \overset{\text{def}}{=} \big(\mathbf{X}^T\mathbf{A}\mathbf{X}\big)^{-1}\mathbf{X}^T\mathbf{A}\mathbf{Y}, \tag{5}$$

where $\mathbf{X} \overset{\text{def}}{=} (X^i_j)_{j=1,\dots,N;\ i=1,\dots,d}$ is the $N \times d$ matrix of observed regressors, $\mathbf{Y} \overset{\text{def}}{=} (Y_1,\dots,Y_N)^T$ is the vector of observed responses, and $\mathbf{A} \overset{\text{def}}{=} \operatorname{diag}(a^m_{1;N},\dots,a^m_{N;N})$ is the diagonal weight matrix for estimation of mth component. (Obviously, $\mathbf{A}$ depends on m, but we do not show it explicitly by a subscript since the number m of the component for which $\mathbf{b}_m$ is estimated will be further fixed.)

Generally speaking, by (4) $a^m_{j;N}$ must be negative for some j, so $\hat{\mathbf{b}}^{(m)}_{;N}$ is not necessarily an argmin of $\hat{J}_{m;N}(\mathbf{b})$. But we will take $\hat{\mathbf{b}}^{(m)}_{;N}$ as an estimate for $\mathbf{b}_m$ and call it a modified least-squares estimate for $\mathbf{b}_m$ in MVC model (MVC-LS estimate).

Let

$$\mathbf{D}^{(k)} \overset{\text{def}}{=} \mathsf{E}\big[\mathbf{X}(O)\mathbf{X}^T(O) \;\big|\; \kappa(O)=k\big]$$

be the matrix of second moments of the regressors for subjects belonging to the kth component. Denote the variance of the kth component's error term by

$$\big(\sigma^{(k)}\big)^2 = \mathsf{E}\big[\big(\varepsilon(O)\big)^2 \;\big|\; \kappa(O)=k\big].$$

(Recall that $\mathsf{E}[(\varepsilon(O)) \mid \kappa(O)=k]=0$). In what follows, we assume that these moments and variances exist for all components.

Theorem 1 (Consistency)**.** *Assume that*

1. $\mathbf{D}^{(k)}$ *and* $(\sigma^{(k)})^2$ *are finite for all* $k=1,\dots,M$.
2. $\mathbf{D}^{(m)}$ *is nonsingular.*
3. *There exists* $C>0$ *such that* $\det \boldsymbol{\Gamma}_N > C$ *for all* N *large enough.*

Then $\hat{\mathbf{b}}^{(m)}_{;N} \xrightarrow{\mathrm{P}} \mathbf{b}^{(m)}$ as $N \to \infty$.

Note. Assumption 3 can be weakened. Applying Theorem 4.2 from [8], we can show that $\hat{\mathbf{b}}^{(m)}_{;N}$ is consistent if the vector $\mathbf{p}^m_N$ is asymptotically linearly independent from the vectors $\mathbf{p}^i_N$, $i \neq m$, as $N \to \infty$. To avoid complexities in this presentation, we do not formulate the strict meaning of this statement.

Denote $D^{ik(s)} \overset{\text{def}}{=} \mathsf{E}\big[X^i(O)X^k(O) \mid \kappa(O)=s\big]$,

$$\mathbf{L}^{ik(s)} \overset{\text{def}}{=} \big(\mathsf{E}\big[X^i(O)X^k(O)X^q(O)X^l(O) \mid \kappa(O)=s\big]\big)_{l,q=1}^{d},$$

$$\mathbf{M}^{ik(s,p)} \overset{\text{def}}{=} \big(D^{il(s)}D^{kq(p)}\big)_{l,q=1}^{d}.$$

Theorem 2 (Asymptotic normality)**.** *Assume that*

1. $\mathsf{E}[(X^i(O))^4 \mid \kappa(O)=k] < \infty$ *and* $\mathsf{E}[(\varepsilon(O))^4 \mid \kappa(O)=k] < \infty$ *for all* $k=1,\dots,M$.
2. *Matrix* $\mathbf{D}=\mathbf{D}^{(m)}$ *is nonsingular.*
3. *There exists* $C>0$ *such that* $\det \boldsymbol{\Gamma}_N > C$ *for all* N *large enough.*
4. *For all* $s,p=1,\dots,M$, *there exist* $\langle(\mathbf{a}^m)^2\mathbf{p}^s\mathbf{p}^p\rangle$.

Then $\sqrt{N}(\hat{\mathbf{b}}^{(m)}_{;N}-\mathbf{b}^{(m)}) \xrightarrow{\text{W}} N(0,\mathbf{V})$, where

$$\mathbf{V} \overset{\text{def}}{=} \mathbf{D}^{-1}\boldsymbol{\Sigma}\mathbf{D}^{-1} \tag{6}$$

with

$$\boldsymbol{\Sigma} = \big(\Sigma^{ik}\big)_{ik=1}^{d},$$

$$\begin{aligned}\Sigma^{ik} &= \sum_{s=1}^{M}\big\langle(\mathbf{a}^m)^2\mathbf{p}^s\big\rangle\big(D^{ik(s)}\big(\sigma^{(s)}\big)^2 + \big(\mathbf{b}^{(s)}-\mathbf{b}^{(m)}\big)^T\mathbf{L}^{ik(s)}\big(\mathbf{b}^{(s)}-\mathbf{b}^{(m)}\big)\big)\\ &\quad -\sum_{s=1}^{M}\sum_{p=1}^{M}\big\langle(\mathbf{a}^m)^2\mathbf{p}^s\mathbf{p}^p\big\rangle\big(\mathbf{b}^{(s)}-\mathbf{b}^{(m)}\big)^T\mathbf{M}^{ik(s,p)}\big(\mathbf{b}^{(p)}-\mathbf{b}^{(m)}\big).\end{aligned} \tag{7}$$

4 Proofs

Proof of Theorem 1. Note that if $\mathbf{D}^{(m)}$ is nonsingular, then

$$\mathbf{b}^{(m)} = \operatorname{argmin}_{\mathbf{b}\in\mathbb{R}^d} J_{m;\infty}(\mathbf{b}) = \big(\mathbf{D}^{(m)}\big)^{-1}\mathsf{E}\big[\big(Y(O)\mathbf{X}(O)\big) \mid \kappa(O)=m\big].$$

By Lemma 1,

$$\mathbf{X}^T\mathbf{A}\mathbf{X} = \frac{1}{N}\sum_{j=1}^{N} a^m_{j;N}\mathbf{X}(O_j)\mathbf{X}^T(O_j) \xrightarrow{\text{P}} \mathbf{D}^{(m)}$$

and

$$\mathbf{X}^T\mathbf{A}\mathbf{Y} = \frac{1}{N}\sum_{j=1}^{N} a^m_{j;N}Y(O_j)\mathbf{X}(O_j) \xrightarrow{\text{P}} \mathsf{E}\big[\big(Y(O)\mathbf{X}(O)\big) \mid \kappa(O)=m\big]$$

as $N\to\infty$. This implies the statement of the theorem. □

Proof of Theorem 2. Let us introduce a set of random vectors $\xi_j^{(k)} = (Y_j^{(k)}, X_j^{1(k)}, \dots, X_j^{d(k)})^T$, $j = 1, 2, \dots$, with distributions F_k that are independent for different j and k and independent from κ_j. Denote $\delta_j^{(k)} = \mathbb{1}\{\kappa_j = k\}$,

$$\xi_j' \overset{\text{def}}{=} \sum_{k=1}^{M} \delta_j^{(k)} \xi_j^{(k)}.$$

Then the distribution of $\varXi_N' = (\xi_1', \dots, \xi_N')$ is the same as that of $\varXi_N$. Since in this theorem we are interested in weak convergence only, without loss of generality, let us assume that $\varXi_N = \varXi_N'$. By $\mathfrak{F}$ we denote the sigma-algebra generated by $\xi_j^{(k)}$, $j = 1, \dots, N$, $k = 1, \dots, M$.

Let us show that $\sqrt{N}(\hat{\mathbf{b}}_{;N}^{(m)} - \mathbf{b}^{(m)})$ converges weakly to $N(0, \mathbf{V})$. It is readily seen that

$$\sqrt{N}\big(\hat{\mathbf{b}}_{;N}^{(m)} - \mathbf{b}^{(m)}\big) = \bigg[\frac{1}{N}\big(\mathbf{X}^T\mathbf{A}\mathbf{X}\big)\bigg]^{-1}\bigg[\frac{1}{\sqrt{N}}\big(\mathbf{X}^T\mathbf{A}\mathbf{Y} - \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{b}^{(m)}\big)\bigg].$$

Since $\frac{1}{N}(\mathbf{X}^T\mathbf{A}\mathbf{X}) \xrightarrow{\text{P}} \mathbf{D}$, we need only to show week convergence of the random vectors $\frac{1}{\sqrt{N}}(\mathbf{X}^T\mathbf{A}\mathbf{Y} - \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{b}^{(m)})$ to $N(0, \boldsymbol{\Sigma})$.

Denote

$$g(\xi_j) \overset{\text{def}}{=} \Bigg(X_j^i\Bigg(Y_j - \sum_{l=1}^{d} X_j^l b_l^{(m)}\Bigg)\Bigg)_{i=1}^{d},$$

$$\gamma_j^i \overset{\text{def}}{=} \frac{1}{\sqrt{N}} a_{j;N}^{(m)} X_j^i\Bigg(Y_j - \sum_{l=1}^{d} X_j^l b_l^{(m)}\Bigg).$$

Obviously,

$$\zeta_N \overset{\text{def}}{=} \frac{1}{\sqrt{N}}\big(\mathbf{X}^T\mathbf{A}\mathbf{Y} - \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{b}^{(m)}\big) = \sqrt{N}\hat{g}_{;N}^{(m)} = \Bigg(\sum_{j=1}^{N}\gamma_j^i\Bigg)_{i=1}^{d}.$$

We will apply Lemma 2 to show that $\sqrt{N}\hat{g}_{;N}^{(m)} \xrightarrow{\text{W}} N(0, \boldsymbol{\Sigma})$.

First, let us show that $\bar{g}^{(m)} = \mathsf{E}\,\hat{g}_{;N}^{(m)} = 0$. It is equivalent to $\mathsf{E}\sum_{j=1}^{N}\gamma_j^i = 0$ for all $i = 1, \dots, d$. In fact,

$$\begin{aligned}
&\mathsf{E}\sum_{j=1}^{N}\gamma_j^i \\
&\quad= \mathsf{E}\Bigg[\mathsf{E}\,\frac{1}{\sqrt{N}}\sum_{j=1}^{N} a_{j;N}^{(m)} \sum_{s=1}^{M}\delta_j^{(s)} X_j^{i(s)}\Bigg(\sum_{s=1}^{M}\delta_j^{(s)} Y_j(s) - \sum_{l=1}^{d}\sum_{s=1}^{M}\delta_j^{(s)} X_j^{l(s)} b_l^{(m)}\Bigg)\Bigg|\mathfrak{F}\Bigg] \\
&\quad= \frac{1}{\sqrt{N}}\,\mathsf{E}\sum_{j=1}^{N} a_{j;N}^{(m)} \sum_{s=1}^{M} p_{j;N}^{s} X_j^{i(s)}\Bigg(Y_j^{(s)} - \sum_{l=1}^{d} X_j^{l(s)} b_l^{(m)}\Bigg)
\end{aligned}$$

$$= \sqrt{N}\sum_{s=1}^{M}\underbrace{\langle a^{(m)}\mathbf{p}^{s}\rangle_N}_{\mathbb{1}\{m=s\}}\mathsf{E}\Bigg[X_1^{i(s)}\Bigg(Y_1^{(s)}-\sum_{l=1}^{d}X_1^{l(s)}b_l^{(m)}\Bigg)\Bigg]$$

$$= \sqrt{N}\,\mathsf{E}\Bigg[X_1^{i(m)}\underbrace{\Bigg(Y_1^{(m)}-\sum_{l=1}^{d}X_1^{l(m)}b_l^{(m)}\Bigg)}_{\varepsilon_1^{(m)}}\Bigg] = \sqrt{N}\,\mathsf{E}\,X_1^{i(m)}\,\mathsf{E}\,\varepsilon_1^{(m)} = 0.$$

So

$$\zeta_N = \Bigg(\sum_{j=1}^{N}\gamma_j^i\Bigg)_{i=1}^{d} = \Bigg(\sum_{j=1}^{N}\gamma_j^i - \mathsf{E}\,\gamma_j^i\Bigg)_{i=1}^{d}.$$

In view of Lemma 2, to complete the proof, we only need to show thatCov$(\zeta_N) \to \boldsymbol{\Sigma}$.

Denote

$$\zeta_j^{i(s)} = \delta_j^{(s)}X_j^{i(s)}\Bigg(\sum_{l=1}^{d}X_j^{l(s)}\big(b_l^{(s)}-b_l^{(m)}\big)+\varepsilon_j^{(s)}\Bigg)$$

and

$$\eta_j^{i(s)} = \delta_j^{(s)}\Bigg(\sum_{l=1}^{d}D^{il(s)}\big(b_l^{(s)}-b_l^{(m)}\big)\Bigg)$$

Then

$$\zeta_N = \Bigg(\frac{1}{\sqrt{N}}\sum_{j=1}^{N}a_j^{(m)}\sum_{s=1}^{M}\big(\big(\zeta_j^{i(s)}-\eta_j^{i(s)}\big)+\big(\eta_j^{i(s)}-\mathsf{E}\,\zeta_j^{i(s)}\big)\big)\Bigg)_{i=1}^{d} = \big(S_1^i+S_2^i\big)_{i=1}^{d},$$

where

$$S_1^i = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}a_j^{(m)}\sum_{s=1}^{M}\big(\zeta_j^{i(s)}-\eta_j^{i(s)}\big), \qquad S_2^i = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}a_j^{(m)}\sum_{s=1}^{M}\big(\eta_j^{i(s)}-\mathsf{E}\,\zeta_j^{i(s)}\big).$$

Note that

$$\mathsf{E}\big(\zeta_j^{i(s)}-\eta_j^{i(s)}\big) = \mathsf{E}\,\mathsf{E}\big[\big(\zeta_j^{i(s)}-\eta_j^{i(s)}\big)\mid\kappa_j\big] = 0.$$

Now

$$\begin{aligned}
&\mathrm{Cov}\big(S_1^i, S_2^k\big)\\
&\quad= \frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}\sum_{p=1}^{M}\mathsf{E}\big(\zeta_j^{i(s)}-\eta_j^{i(s)}\big)\eta_j^{k(p)}\\
&\quad= \frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}\sum_{p=1}^{M}\mathsf{E}\,\delta_j^{(s)}\Bigg[\sum_{l=1}^{d}\big(X_j^{i(s)}X_j^{l(s)}-D^{il(s)}\big)\big(b_l^{(s)}-b_l^{(m)}\big)\\
&\qquad+X_j^{i(s)}\varepsilon_j^{i(s)}\Bigg]\cdot\delta_j^{(p)}\Bigg(\sum_{q=1}^{d}D^{kq(p)}\big(b_q^{(p)}-b_q^{(m)}\big)\Bigg)
\end{aligned}$$

$$= \frac{1}{N}\sum_{j=1}^{N}\big(a_j^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^s\Bigg(\sum_{l=1}^{d}\big(\underbrace{D^{il(s)}-D^{il(s)}}_{0}\big)\big(b_l^{(s)}-b_l^{(m)}\big)+\mathsf{E}\,X_j^{i(s)}\underbrace{\mathsf{E}\,\varepsilon_j^{(s)}}_{0}\Bigg)$$

$$\times\Bigg(\sum_{q=1}^{d}D^{kq(s)}\big(b_q^{(s)}-b_q^{(m)}\big)\Bigg)=0; \tag{8}$$

$$\begin{aligned}\operatorname{Cov}\big(S_1^i,S_1^k\big)&=\frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^s\sum_{q=1}^{d}\sum_{l=1}^{d}\big(b_l^{(s)}-b_l^{(m)}\big)\,\mathsf{E}\big(\big(X_j^{i(s)}X_j^{l(s)}-D^{il(s)}\big)\\&\quad\times\big(X_j^{k(s)}X_j^{q(s)}-D^{kq(s)}\big)\big)\big(b_q^{(s)}-b_q^{(m)}\big)\\&\quad+\frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^sD^{ik(s)})\big(\sigma^{(s)}\big)^2;\end{aligned}$$

$$\begin{aligned}\operatorname{Cov}\big(S_2^i,S_2^k\big)&=\frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^s\sum_{q=1}^{d}\sum_{l=1}^{d}\big(b_l^{(s)}-b_l^{(m)}\big)D^{il(s)}\\&\quad\times\Bigg[\big(b_q^{(s)}-b_q^{(m)}\big)D^{kq(s)}-\sum_{r=1}^{M}p_{j;N}^r\big(b_q^{(r)}-b_q^{(m)}\big)D^{kq(r)}\Bigg].\end{aligned}$$

Thus,

$$\begin{aligned}\operatorname{Cov}\zeta_N&=\frac{1}{N}\sum_{j=1}^{N}\big(a_{j;N}^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^sD^{ik(s)}\big(\sigma^{(s)}\big)^2\\&\quad+\frac{1}{N}\sum_{j=1}^{N}\big(a_j^{(m)}\big)^2\sum_{s=1}^{M}p_{j;N}^s\sum_{q=1}^{d}\sum_{l=1}^{d}\big(b_l^{(s)}-b_l^{(m)}\big)\Bigg[\big(b_q^{(s)}-b_q^{(m)}\big)\\&\quad\times\underbrace{\mathsf{E}\,X_j^{i(s)}X_j^{k(s)}X_j^{l(s)}X_j^{q(s)}}_{L_{lq}^{ik(s)}}-D^{il(s)}\sum_{r=1}^{M}p_{j;N}^r\big(b_q^{(r)}-b_q^{(m)}\big)D^{kq(r)}\Bigg]\Bigg)_{i,k=1}^{d}.\end{aligned}$$

From the last equation we get $\operatorname{Cov}(\zeta_N)\to\boldsymbol{\Sigma}$ as $N\to\infty$. □

5 Results of simulation

To assess the accuracy of the asymptotic results from Section 3, we performed a small simulation study. We considered a two-component mixture ($M=2$) with mixing probabilities $p_{j;N}^1=j/N$ and $p_{j;N}^2=1-p_{j;N}^1$. For each subject, there were two observed variables X and Y, which were simulated based on the simple linear

Table 1. Simulation results

First component					
n	$\mathsf{E}\,\hat{b}_0^1$	$\mathsf{E}\,\hat{b}_1^1$	$N\,\mathrm{Var}\,\hat{b}_0^1$	$N\,\mathrm{Var}\,\hat{b}_1^1$	$N\,\mathrm{Cov}(\hat{b}_0^1, \hat{b}_1^1)$
500	3.0014	0.5235	47.61	45.83	– 42.27
1000	3.0033	0.512	41.19	37.50	−35.04
2000	3.0011	0.5032	38.84	34.71	−32.87
5000	3.0003	0.5016	39.54	34.49	−32.83
∞	3	0.5	39.13	33.96	−32.53
Second component					
n	$\mathsf{E}\,\hat{b}_0^2$	$\mathsf{E}\,\hat{b}_1^2$	$N\,\mathrm{Var}\,\hat{b}_0^2$	$N\,\mathrm{Var}\,\hat{b}_1^2$	$N\,\mathrm{Cov}(\hat{b}_0^2, \hat{b}_1^2)$
500	−2.0243	1.0084	67.37	7.94	−22.17
1000	−42.0100	1.0027	63.04	7.57	−20.90
2000	−2.0039	1.0016	63.57	7.52	−20.95
5000	−2.0074	1.0025	62.41	7.32	−20.48
∞	−2	1	62.20	7.34	−20.47

regression model

$$Y_j = b_0^{\kappa_j} + b_1^{\kappa_j} X_j^{(\kappa_j)} + \varepsilon_j^{(\kappa_j)},$$

where κ_i is the number of component the jth observation belongs to, $X_j^{(1)}$ was simulated as $N(1, 1)$, $X_j^{(2)}$ as $N(2, 2.25)$, and $\varepsilon_j^{(k)}$ were zero-mean Gaussians with standard deviations 0.01 for the first component and 0.05 for the second one. The values of the regression coefficients were $b_0^1 = 3$, $b_1^1 = 0.5$, $b_0^2 = -2$, $b_1^2 = 1$.

The means and covariances of the estimates were calculated over 2000 replications.

The results of simulation are presented in Table 1. The true values of parameters and asymptotic covariances are placed in the last rows of the tables.

The presented data show good concordance with the asymptotic theory for $n > 1000$.

6 Conclusions

We considered a modification of least-squares estimators for linear regression coefficients in the case where observations are obtained from a mixture with varying concentrations. Conditions of consistency and asymptotic normality of the estimators were derived, and dispersion matrices were evaluated. The results of simulations confirm good concordance of estimators covariances with the asymptotic formulas for sample sizes larger then 1000 observations.

In real-life data analysis, concentrations (mixing probabilities) are usually not known exactly but estimated. So, to apply the proposed technique, we also need to analyze sensitivity of the estimates to perturbations of the concentrations model. (We are thankful to the unknown referee for this observation). It is worth noting that performance of these estimates will be poor if the true concentrations of the components are nearly linearly dependent ($\det \boldsymbol{\Gamma}_N \approx 0$). We also expect stability of the estimates w.r.t. concentration perturbations if $\det \boldsymbol{\Gamma}_N$ is bounded away from zero. More deep analysis of sensitivity will be a part of our further work.

References

[1] Autin, F., Pouet, Ch.: Test on the components of mixture densities. Stat. Risk. Model. **28**(4), 389–410 (2011). MR2877572. doi:10.1524/strm.2011.1065

[2] Billingsley, P.: Probability and Measure, 3rd edn. Wiley, New York (1995). MR1324786

[3] Borovkov, A.A.: Mathematical Statistics. Gordon and Breach Science Publishers, Amsterdam (1998). MR1712750

[4] Doronin, O.V.: Lower bound for a dispersion matrix for the semiparametric estimation in a model of mixtures. Theory Probab. Math. Stat. **90**, 71–85 (2015). MR3241861. doi:10.1090/tpms/950

[5] Grün, B., Leisch, F.: Fitting finite mixtures of linear regression models with varying & fixed effects in R. In: Rizzi, A., Vichi, M. (eds.) Compstat 2006 – Proceedings in Computational Statistics, pp. 853–860. Physica Verlag, Heidelberg, Germany (2006)

[6] Maiboroda, R.: Statistical Analysis of Mixtures. Kyiv University Publishers, Kyiv (2003) (in Ukrainian)

[7] Maiboroda, R.E., Sugakova, O.V.: Estimation and classification by observations from mixture. Kyiv University Publishers, Kyiv (2008) (in Ukrainian)

[8] Maiboroda, R., Sugakova, O.: Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. J. Nonparametr. Stat. **24**(1), 201–205 (2012). MR2885834. doi:10.1080/10485252.2011.630076

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

[10] McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley–Interscience, New York (2000). MR1789474. doi:10.1002/0471721182

[11] Seber, G.A.F., Lee, A.J.: Linear Regression Analysis. Wiley, New York (2003). MR1958247. doi:10.1002/9780471722199

[12] Shcherbina, A.: Estimation of the mean value in the model of mixtures with varying concentrations. Theory Probab. Math. Stat. **84**, 151–164 (2012). MR2857425. doi:10.1090/S0094-9000-2012-00866-1

[13] Titterington, D.M., Smith, A.F., Makov, U.E.: Analysis of Finite Mixture Distributions. Wiley, New York (1985)