Modern Stochastics: Theory and Applications 5 (2) (2018) 225–245
https://doi.org/10.15559/18-VMSTA105

Confidence ellipsoids for regression coefficients by observations from a mixture

Vitalii Miroshnichenko, Rostyslav Maiboroda*

Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

vitaliy.miroshnychenko@gmail.com (V. Miroshnichenko), mre@univ.kiev.ua (R. Maiboroda)

Received: 29 January 2018, Revised: 16 May 2018, Accepted: 19 May 2018,
Published online: 4 June 2018

Abstract Confidence ellipsoids for linear regression coefficients are constructed by observations from a mixture with varying concentrations. Two approaches are discussed. The first one is the nonparametric approach based on the weighted least squares technique. The second one is an approximate maximum likelihood estimation with application of the EM-algorithm for the estimates calculation.

Keywords Finite mixture model, linear regression, mixture with varying concentrations, nonparametric estimation, maximum likelihood, confidence ellipsoid, EM-algorithm
2010 MSC 62J05, 62G20

1 Introduction

This paper is devoted to the technique of construction of confidence ellipsoids for coefficients of linear regression in the case, when statistical data are derived from a mixture with finite number of components and the mixing probabilities (the concentrations of components) are different for different observations. These mixing probabilities are assumed to be known, but the distributions of the components are unknown. (Such mixture models are also known as mixtures with varying concentrations, see [1] and [7]).

The problem of estimation of regression coefficients by such mixed observations was considered in the parametric setting in [4] and [5]. The authors of these papers

*Corresponding author.

assume that the distributions of regression errors and regressors are known up to some unknown parameters. The models considered in these papers are called *finite mixtures of regression models*. Some versions of maximum likelihood estimates are used for the estimation of unknown parameters of distributions and regression coefficients under these models. The EM-algorithm is used for computation of the estimates. (This algorithm is also implemented in R package `mixtools`, [2]. See [13] for the general theory of EM-algorithm and its application to mixture models).

In [6] a nonparametric approach was proposed under which no parametric models on the distributions of components are assumed. A weighted least squares technique is used to derive estimates for regression coefficients. Consistency and asymptotic normality of the estimates are demonstrated.

Note that in [6, 10, 11] a nonparametric approach to the analysis of mixtures with varying concentrations was developed in the case when the concentrations of components (mixing probabilities) are known. Some examples of real life data analysis under this assumption were considered in [10, 11].

Namely, in [11] an application to the analysis of the Ukrainian parliamentary elections-2006 was considered. Here the observed subjects were respondents of the Four Wave World Values survey (conduced in Ukraine in 2006) and the mixture components were the populations of different electoral behavior adherents. The mixing probabilities were obtained from the official results of voting in 27 regions of Ukraine.

In [10] an application to DNA microarray data analysis was presented. Here the subjects were nearly 3000 of genes of the human genome. They were divided into two components by the difference of their expression in two types of malignantly transformed tissues. The concentrations were defined as a posteriori probabilities for the genes to belong to a given component, calculated by observations on the genes expression in sample tissues.

In this paper we will show how to construct confidence sets (ellipsoids) for regression coefficients under both parametric and nonparametric approaches. Quality of obtained ellipsoids is compared via simulations.

The rest of the paper in organized as follows. In Section 2 we present a formal description of the model. Nonparametric and parametric estimates of regression coefficients and their asymptotic properties are discussed in Sections 3 and 4. Estimation of asymptotic covariances of these estimates is considered in Section 5. The confidence ellipsoids are constructed in Section 6. Results of simulations are presented in Section 7. In Section 8 we present a toy example of application to a real life sociological data. Section 9 contains concluding remarks.

2 The model

We consider the model of mixture with varying concentrations. It means that each observed subject O belongs to one of M different subpopulations (mixture components). The number of component which the subject belongs to is denoted by $\kappa(O) \in \{1, 2, \dots, M\}$. This characteristic of the subject is not observed. The vector of observed variables of O will be denoted by $\xi(O)$. It is considered as a random vector with the distribution depending on the subpopulation which O belongs to.

A structural linear regression model will be used to describe these distributions. (See [14] for general theory of linear regression).

That is, we consider one variable $Y = Y(O)$ in $\xi(O) = (Y(O), X^1(O), \dots, X^d(O))^T$ as a response and all other ones $\mathbf{X}(O) = (X^1(O), \dots, X^d(O))^T$ as regressors in the model

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

where b_i^k, $i = 1, \dots, d$, $k = 1, \dots, M$ are unknown regression coefficients for the k-th component of the mixture, $\varepsilon(O)$ is the error term. Denote by $\mathbf{b}^k = (b_1^k, \dots, b_d^k)^T$ the vector of the k-th component's coefficients. We consider $\varepsilon(O)$ as a random variable and assume that

$$\mathsf{E}\big[\varepsilon(O) \mid \kappa(O) = m\big] = 0, \quad m = 1, \dots, M,$$

and

$$\sigma_m^2 = \mathrm{Var}\big[\varepsilon(O) \mid \kappa(O) = m\big] < \infty.$$

(σ_m^2 are unknown).

It is also assumed that the regression error term $\varepsilon(O)$ and regressors $\mathbf{X}(O)$ are conditionally independent for fixed $\kappa(O) = m$, $m = 1, \dots, M$.

The observed sample Ξ_n consists of values $\xi_j = (Y_j, \mathbf{X}_j^T)^T = \xi(O_j)$, $j = 1, \dots, n$, where $O_1, \dots, O_n$ are independent subjects which can belong to different components with probabilities

$$p_j^m = \mathsf{P}\big\{\kappa(O_j) = m\big\}, \quad m = 1, \dots, M; \ j = 1, \dots, n.$$

(all mixing probabilities p_j^m are known).

To describe completely the probabilistic behavior of the observed data we need to introduce the distributions of $\varepsilon(O)$ and $\mathbf{X}(O)$ for different components. Let us denote

$$F_{\mathbf{X},m}(A) = \mathsf{P}\big\{\mathbf{X}(O) \in A \mid \kappa(O) = m\big\} \text{ for any measurable } A \subseteq \mathbb{R}^d,$$

and

$$F_{\varepsilon,m}(A) = \mathsf{P}\big\{\varepsilon(O) \in A \mid \kappa(O) = m\big\} \text{ for any measurable } A \subseteq \mathbb{R}.$$

The corresponding probability densities $f_{\mathbf{X},m}$ and $f_{\varepsilon,m}(x)$ are defined by

$$F_{\mathbf{X},m}(A) = \int_A f_{\mathbf{X},m}(\mathbf{x})d\mathbf{x}, \ \ F_{\varepsilon,m}(A) = \int_A f_{\varepsilon,m}(x)dx$$

(for all measurable A).

The distribution of observed ξ_j is a mixture of distributions of components with the mixing probabilities p_j^m, e.g.

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

and the probability density $f_j(y, \mathbf{x})$ of $\xi_j = (Y_j, \mathbf{X}_j^T)^T$ at a point $(y, \mathbf{x}^T)^T \in \mathbb{R}^{d+1}$ is

$$f_j(y, \mathbf{x}) = \sum_{m=1}^{M} p_j^m f_{\mathbf{X},m}(\mathbf{x}) f_{\varepsilon,m}\big(y - \mathbf{x}^T \mathbf{b}^m\big).$$

In what follows we will discuss two approaches to the estimation of the parameters of interest $\mathbf{b}^k$, for a fixed $k \in \{1, \dots, M\}$.

The first one is the nonparametric approach. Under this approach we do not need to know the densities $f_{\mathbf{X},m}$ and $f_{\varepsilon,m}$. Moreover we even do not assume the existence of these densities. The estimates are based on some modification of the least squares tehnique proposed in [6].

In the second, parametric approach we assume that the densities of components are known up to some unknown nuisance parameters $\vartheta_m \in \Theta \subseteq \mathbb{R}^L$:

$$f_{\mathbf{X},m}(\mathbf{x}) = f(\mathbf{x}; \vartheta_m), \qquad f_{\varepsilon,m}(x) = f_\varepsilon(x; \vartheta_m). \tag{2}$$

In the most popular parametric *normal mixture model* these densities are normal, i.e.

$$f_{\varepsilon,m} \sim N\big(0, \sigma_m^2\big), \qquad f_{\mathbf{X},m}(\mathbf{x}) \sim N(\mu_m, \Sigma_m), \tag{3}$$

where $\mu_m \in \mathbb{R}^d$ is the mean of $\mathbf{X}$ for the m-th component and $\Sigma_m \in \mathbb{R}^{d\times d}$ is its covariance matrix. All the parameters are usually unknown. So, in this case the unknown nusance parameters are

$$\vartheta_m = \big(\mu_m, \Sigma_m, \sigma_m^2\big), \quad m = 1, \dots, M.$$

3 Generalized least squares estimator

Let us consider the nonparametric approach to the estimation of the regression coefficients developed in [6]. It is based on the minimization of weighted least squares

$$J_{k;n}(\mathbf{b}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{j=1}^{n} a_{j;n}^k \left(Y_j - \sum_{i=1}^{d} b_i X_j^i \right)^2,$$

over all possible $\mathbf{b} = (b_1, \dots, b_d)^T \in \mathbb{R}^d$.

Here $\mathbf{a}^k = (a_{1;n}^k, \dots, a_{n;n}^k)$ are the minimax weights for estimation of the k-th component's distribution. They are defined by

$$a_{j;n}^k = \frac{1}{\det \boldsymbol{\Gamma}_n} \sum_{m=1}^{M} (-1)^{k+m} \gamma_{mk;n} p_j^m, \tag{4}$$

where $\gamma_{mk;n}$ is the (mk)-th minor of the matrix

$$\boldsymbol{\Gamma}_n = \left(\frac{1}{n} \sum_{j=1}^{n} p_j^l p_j^i \right)_{l,i=1}^{M},$$

see [6, 10] for details.

Define $\mathbf{X} \stackrel{\text{def}}{=} (X_j^i)_{j=1,\dots,n;\ i=1,\dots,d}$ to be the $n \times d$-matrix of observed regressors, $\mathbf{Y} \stackrel{\text{def}}{=} (Y_1, \dots, Y_N)^T$ be the vector of observed responses, $\mathbf{A} \stackrel{\text{def}}{=} \operatorname{diag}(a_{1;n}^k, \dots, a_{n;n}^k)$

be the diagonal weights matrix for estimation of k-th component. Then the stationarity condition

$$\frac{\partial J_{k;n}(\mathbf{b})}{\partial \mathbf{b}} = 0$$

has the unique solution in $\mathbf{b}$,

$$\hat{\mathbf{b}}^{LS}(k,n) \stackrel{\text{def}}{=} \left(\mathbf{X}^T\mathbf{A}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{A}\mathbf{Y}, \tag{5}$$

if the matrix $\mathbf{X}^T\mathbf{A}\mathbf{X}$ is nonsingular.

Note that the weight vector $\mathbf{a}^k$ defined by (4) contains negative weights, so $\hat{\mathbf{b}}_{LS}(k,n)$ is not allways the point of minimum of $J_{k;n}(\mathbf{b})$. But in what follows we will consider $\hat{\mathbf{b}}_{LS}(k,n)$ as a generalized least squares estimate for $\mathbf{b}^k$.

The asymptotic behavior of $\hat{\mathbf{b}}_{LS}(k,n)$ as $n \to \infty$ was investigated in [6]. To describe it we will need some additional notation.

Let us denote by

$$\mathbf{D}^{(m)} \stackrel{\text{def}}{=} \mathsf{E}\left[\mathbf{X}(O)\mathbf{X}^T(O) \mid \kappa(O) = m\right]$$

the matrix of second moments of regressors for the m-th component.

The consistency conditions for the estimator $\hat{\mathbf{b}}^{LS}(k,n)$ are given by the following theorem.

Theorem 1 (Theorem 1 in [6])**.** *Assume that*

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

Then $\hat{\mathbf{b}}^{LS}(k,n) \xrightarrow{\text{P}} \mathbf{b}^{(k)}$ *as* $n \to \infty$.

Let us introduce the following notation.

$$\begin{aligned}
D^{is(m)} &\stackrel{\text{def}}{=} \mathsf{E}\left[X^i(O)X^s(O)|\kappa(O) = m\right],\\
\mathbf{L}^{is(m)} &\stackrel{\text{def}}{=} \left(\mathsf{E}\left[X^i(O)X^s(O)X^q(O)X^l(O)|\kappa(O) = m\right]\right)_{l,q=1}^{d},\\
\mathbf{M}^{is(m,p)} &\stackrel{\text{def}}{=} \left(D^{il(m)}D^{sq(p)}\right)_{l,q=1}^{d},\\
\alpha_{s,q}^{(k)} &= \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\left(a_{j;n}^k\right)^2 p_j^s p_j^q
\end{aligned} \tag{6}$$

(if this limit exists),

$$\alpha_s^{(k)} = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\left(a_{j;n}^k\right)^2 p_j^s = \sum_{q=1}^{M}\alpha_{s,q}^{(k)}.$$

The following theorem provides conditions for the asymptotic normality and describes the dispersion matrix of the estimator $\hat{\mathbf{b}}^{LS}(k,n)$.

Theorem 2 (Theorem 2 in [6]). *Assume that*

1. $\mathsf{E}[(X^i(O))^4 \mid \kappa(O) = m] < \infty$ *and* $\mathsf{E}[(\varepsilon(O))^4 \mid \kappa(O) = m] < \infty$ *for all* $m = 1, \dots, M$, $i = 1, \dots, d$.
2. *Matrix* $\mathbf{D} = \mathbf{D}^{(k)}$ *is nonsingular.*
3. *There exists* $C > 0$ *such that* $\det \boldsymbol{\Gamma}_n > C$ *for all n large enough.*
4. *For all* $s,q = 1, \dots, M$ *there exist* $\alpha_{s,q}^{(k)}$ *defined by (6).*

Then $\sqrt{n}(\hat{\mathbf{b}}^{LS}(k,n) - \mathbf{b}^{(k)}) \xrightarrow{\mathrm{W}} N(0, \mathbf{V})$, *where*

$$\mathbf{V} \stackrel{\mathrm{def}}{=} \mathbf{D}^{-1} \boldsymbol{\Sigma} \mathbf{D}^{-1}, \tag{7}$$

$$\boldsymbol{\Sigma} = \left(\Sigma^{il}\right)_{il=1}^{d},$$

$$\Sigma^{il} = \sum_{s=1}^{M} \alpha_s^k \left(D^{il(s)} \sigma_s^2 + \left(\mathbf{b}^s - \mathbf{b}^k\right)^T \mathbf{L}^{il(s)} \left(\mathbf{b}^s - \mathbf{b}^k\right)\right)$$

$$- \sum_{s=1}^{M} \sum_{m=1}^{M} \alpha_{s,m}^k \left(\mathbf{b}^s - \mathbf{b}^k\right)^T \mathbf{M}^{il(s,m)} \left(\mathbf{b}^m - \mathbf{b}^k\right). \tag{8}$$

4 Parametric approach

In this section we discuss the parametric approach to the estimation of $\mathbf{b}^k$ based on papers [4, 5]. We will assume that the representation (2) holds with some unknown $\vartheta_m \in \Theta \subseteq \mathbb{R}^L$, $m = 1, \dots, M$.

Then the set of all unknown parameters $\tau = (\mathbf{b}^k, \beta, \vartheta)$ consists of

$$\beta = \left(\mathbf{b}^m, m = 1, \dots, M, m \neq k\right)$$

and

$$\vartheta = (\vartheta_1, \dots, \vartheta_M).$$

Here $\mathbf{b}^k$ is our parameter of interest, β and ϑ are the nuisance parameters.

In this model the log-likelihood for the unknown τ by the sample Ξ_n can be defined as

$$L(\tau) = \sum_{j=1}^{n} L(\xi_j, \mathbf{p}_j, \tau),$$

where $\mathbf{p}_j = (p_j^1, \dots, p_j^M)^T$,

$$L(\xi_j, \mathbf{p}_j, \tau) = \ln\left(\sum_{m=1}^{M} p_j^m f_{X,m}(\mathbf{X}_j; \vartheta_m) f_{\varepsilon,m}\left(\mathbf{Y}_j - \mathbf{X}_j^T \mathbf{b}^m\right)\right).$$

The general maximum likelihood estimator (MLE) $\hat{\tau}_n^{MLE} = (\hat{\mathbf{b}}^{k,MLE}, \hat{\beta}^{MLE}, \hat{\vartheta}^{MLE})$ for τ is defined as

$$\hat{\tau}_n^{MLE} = \mathrm{argmax}_\tau \, L(\tau),$$

where the maximum is taken over all possible values of τ. Unfortunately, this estimator is not applicable to most common parametric mixture models, since the log-likelihood $L(\tau)$ usually is not bounded on the set of all possible τ.

For example, it is so in the normal mixture model (3). Really, in this model $L(\tau) \to \infty$ as $\sigma_1^2 \to 0$ and $Y_1 - \mathbf{X}_1^T \mathbf{b}^1 = 0$ with all other parameters being arbitrary fixed.

The usual way to cope with this problem is to use the one-step MLE, which can be considered as one iteration of the Newton–Raphson algorithm of approximate calculation of MLE, starting from some pilot estimate (see [15], section 4.5.3). Namely, let $\hat{\tau}_n^{(0)}$ be some pilot estimate for τ. Let us consider τ as a vector of dimension $P = d \times M + M \times L$ and denote its entries by τ_i:

$$\tau = (\tau_1, \ldots, \tau_P).$$

Denote the gradient of $L(\tau)$ by

$$s_n(\tau) = \frac{\partial L(\tau)}{\partial \tau} = \left(\frac{\partial L(\tau)}{\partial \tau_1}, \ldots, \frac{\partial L(\tau)}{\partial \tau_P} \right)^T$$

and the Hessian of $L(\tau)$ by

$$\gamma_n(\tau) = \left(\frac{\partial L(\tau)}{\partial \tau_i \tau_l} \right)_{i,l=1}^{P}.$$

Then the one-step estimator for τ starting from $\hat{\tau^{(0)}}$ is defined as

$$\hat{\tau}_n^{OS} = \hat{\tau}^{(0)} - \left(\gamma_n\left(\hat{\tau}^{(0)}\right) \right)^{-1} s_n\left(\hat{\tau}^{(0)}\right).$$

Theorem 4.19 in [15] provides general conditions under which $\hat{\tau}_n^{OS}$ constructed by an i.i.d. sample is consistent, asymptotically normal and asymptotically efficient.[1] The limit distribution of the normalized one-step estimate is the same as of the consistent version of MLE.

So, if the assumptions of theorem 4.19 (or other analogous statement) hold, there is no need to use an iterative procedure to derive an estimate with asymptotically optimal performance. But on samples of moderate size $\hat{\tau}_n^{OS}$ can be not good enough.

Another popular way to obtain a stable estimate for τ is to use some version of EM-algorithm. A general EM-algorithm is an iterative procedure for approximate calculation of maximum likelihood estimates when information on some variables is missed. We describe here only the algorithm which calculates EM estimates $\hat{\tau}_n^{EM}$ under the normal mixture model assumptions (3), cf. [4, 5].

The algorithm starts from some pilot estimate

$$\hat{\tau}^{(0)} = \left(\hat{\mathbf{b}}_m^{(0)}, \hat{\sigma}_m^{2(0)}, \hat{\mu}_m^{(0)}, \hat{\Sigma}_m^{(0)}, m = 1, \ldots, M \right)$$

for the full set of the model parameters.

[1] Note that in our setting the sample Ξ_n is not an i.i.d. sample. But one can consider it as i.i.d. if the vectors of concentrations $(p_j^1, \ldots, p_j^M)$ are generated by some stochastic mechanism as i.i.d. vectors. See [12] for an example of such *stochastic concentrations* models.

Then for $i = 1, 2, \dots$ the estimates are iteratively recalculated in the following way.

Assume that on the i-iteration estimates $\hat{\mathbf{b}}^{(i)}$, $\hat{\sigma}_m^{2(i)}$, $\hat{\mu}_m^{(i)}$, $\hat{\Sigma}_m^{(i)}$, $m = 1, \dots, M$ are obtained. Then the i-th stage weights are defined as

$$w_j^{m(i)} = w_j^m\big(\xi_j, \hat{\tau}^{(i)}\big) = \frac{p_j^m f_{\mathbf{X},m}(\mathbf{X}_j; \hat{\mu}_m^{(i)}, \hat{\Sigma}_m^{(i)}) f_{\varepsilon,m}(Y_j - \mathbf{X}_j^T \hat{\mathbf{b}}^{m(i)}; \hat{\sigma}_m^{2(i)})}{\sum_{l=1}^M p_j^l f_{\mathbf{X},l}(\mathbf{X}_j; \hat{\mu}_l^{(i)}, \hat{\Sigma}_l^{(i)}) f_{\varepsilon,l}(Y_j - \mathbf{X}_j^T \hat{\mathbf{b}}^{l(i)}; \hat{\sigma}_l^{2(i)})} \quad (9)$$

(note that $w_j^{m(i)}$ is the posterior probability $\mathsf{P}\{\kappa_j = m \mid \xi_j\}$ calculated for $\tau = \hat{\tau}^{(i)}$).

Let $\bar{w}^m = \sum_{j=1}^n w_j^{m(i)}$. Then the estimators of the $i+1$ iteration are defined as

$$\begin{aligned}
\hat{\mu}_m^{(i+1)} &= \frac{1}{\bar{w}^m} \sum_{j=1}^n w_j^{m(i)} \mathbf{X}_j, \\
\hat{\Sigma}_m^{(i+1)} &= \frac{1}{\bar{w}^m} \sum_{j=1}^n w_j^{m(i)} \big(\mathbf{X}_j - \hat{\mu}_m^{(i)}\big)\big(\mathbf{X}_j - \hat{\mu}_m^{(i)}\big)^T, \\
\hat{\mathbf{b}}^{m(i+1)} &= \left(\sum_{j=1}^n w_j^{m(i)} \mathbf{X}_j \mathbf{X}_j^T \right)^{-1} \sum_{j=1}^n w_j^{m(i)} Y_j \mathbf{X}_j, \\
\hat{\sigma}_m^{2(i+1)} &= \frac{1}{\bar{w}^m} \sum_{j=1}^n w_j^{m(i)} \big(Y_j - \mathbf{X}_j^T \hat{b}^{m(i)}\big)^2.
\end{aligned}$$

The iterations are stopped when some stopping condition is fulfilled. For example, it can be

$$\big\| \hat{\tau}^{(i+1)} - \hat{\tau}^{(i)} \big\| < \delta,$$

where δ is a prescribed target accuracy.

It is known that this procedure provide stable estimates which (for sample large enough) converge to the point of local minimum of $L(\tau)$ which is the closest to the pilot estimator $\hat{\tau}^{(0)}$.

So, this estimator can be considered as an approximate version of a root of likelihood equation estimator (RLE).

The asymptotic behavior of $\hat{\tau}_n^{OS}$ and $\hat{\tau}_n^{EM}$ can be described in terms of Fisher's information matrix $\mathbf{I}^*(n, \tau) = (I_{il}^*(n, \tau))_{i,l=1}^P$, where

$$I_{il}^*(n, \tau) = \sum_{j=1}^n I_{il}(\mathbf{p}_j, \tau), \;\; I_{il}(\mathbf{p}, \tau) = \mathsf{E} \, \frac{\partial L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)}{\partial \tau_i} \frac{\partial L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)}{\partial \tau_l},$$

where $\mathbf{p} = (p^1, \dots, p^m)$, $\xi_{\mathbf{p}}$ is a random vector with the pdf

$$f_{\mathbf{p}}(y, \mathbf{x}; \tau) = \sum_{m=1}^M p^m f(\mathbf{x}; \vartheta_m) f_\varepsilon\big(y - \mathbf{x}^T \mathbf{b}^m; \vartheta_m\big). \quad (10)$$

Under the regularity assumptions (RR) of theorem 70.5 in [3],

$$\big(\mathbf{I}^*(n,\tau)\big)^{1/2}\big(\hat{\tau}_n^{MLE}-\tau\big) \xrightarrow{\mathrm{W}} N(0,\mathbb{E}), \tag{11}$$

where $\mathbb{E}$ is the $R\times R$ unit matrix.

Assumptions (RR) include the assumption of likelihood boundedness, so they do not hold for the normal mixture model. But if the pilot estimate $\hat{\tau}_n^{(0)}$ is $\sqrt{n}$-consistent, one needs only a local version of (RR) to derive asymptotic normality of $\hat{\tau}_n^{OS}$ and $\hat{\tau}_n^{EM}$, i.e. (RR) must hold in some neighborhood of the true value of estimated τ. These local (RR) hold for the normal mixture model if $\sigma_m^2>0$ and Σ_m are nonsingular for all $m=1,\dots,M$.

To use all these results for construction of an estimator for $\mathbf{b}^k$ we need $\sqrt{n}$-consistent pilot estimators for the parameter of interest and nuisance parameters. They can be derived by the nonparametric technique considered in Section 3. To construct confidence ellipsoids we will also need estimators for the dispersion matrix $\mathbf{V}$ from (7) in the nonparametric case and estimators for the information matrix $\mathbf{I}^*(n,\tau)$ in the parametric case. These estimators are discussed in the next section.

5 Estimators for nuisance parameters and normalizing matrices

Let us start with the estimation of the dispersion matrix $\mathbf{V}$ in Theorem 2. In fact, we need to estimate consistently the matrices $\mathbf{D}$ and $\boldsymbol{\Sigma}$.

Note that $\mathbf{D}=\mathbf{D}^{(k)}=(D^{is(k)})_{i,s=1}^d$, where

$$D^{is(k)}=\mathsf{E}\big[X^i(O)X^s(O)\mid\kappa(O)=k\big].$$

By theorem 4.2 in [10],

$$\hat{D}_n^{is(k)}=\frac{1}{n}\sum_{j=1}^n a_j^k X_j^i X_j^s \tag{12}$$

is a consistent estimate for $D^{is(k)}$ if $\mathsf{E}[\|\mathbf{X}(O)\|^2\mid\kappa(O)=m]<\infty$ for all $m=1,\dots,M$ and assumption 3 of Theorem 1 holds.

So one can use $\hat{\mathbf{D}}_n^{(k)}=(\hat{D}_n^{is(k)})_{i,s=1}^d$ as a consistent estimate for $\mathbf{D}$ if the assumptions of Theorem 1 hold.

Similarly, $\mathbf{L}^{is(m)}$ can be estimated consistently by

$$\hat{\mathbf{L}}_n^{is(m)}=\frac{1}{n}\sum_{j=1}^n a_j^m X_j^i X_j^s \mathbf{X}_j\mathbf{X}_j^T \tag{13}$$

under the assumptions of Theorem 2.

The same idea can be used to estimate σ_m^2 by

$$\hat{\sigma}_{m;n}^{2(0)}=\frac{1}{n}\sum_{j=1}^n a_j^m\big(Y_j-\mathbf{X}_j^T\hat{\mathbf{b}}^{LS}(s,n)\big)^2. \tag{14}$$

The coefficients $\alpha_{s,q}^{(k)}$ can be approximated by

$$\hat{\alpha}_{s,q}^{(k)} = \frac{1}{n}\sum_{j=1}^{n}\left(a_{j;n}^{k}\right)^2 p_j^s p_j^q.$$

Now replacing true $\mathbf{D}^{(m)}$, $\mathbf{L}^{is(m)}$, $\mathbf{b}^m$, σ_m^2 and $\alpha_{s,q}^{(k)}$ in formula (8) by their estimators $\hat{\mathbf{D}}_n^{(m)}$, $\hat{\mathbf{L}}_n^{is(m)}$, $\hat{\mathbf{b}}^{LS}(m,n)$, $\hat{\sigma}_{m,n}^2$ and $\hat{\alpha}_{s,q}^{(k)}$, one obtains a consistent estimator $\hat{\Sigma}_n$ for Σ.

Then

$$\hat{\mathbf{V}}_n = \hat{\mathbf{D}}_n^{-1}\hat{\Sigma}_n\hat{\mathbf{D}}_n^{-1} \tag{15}$$

is a consistent estimator for $\mathbf{V}$.

To get the pilot estimators for the normal mixture model one can use the same approach. Namely, we define

$$\hat{\mu}_{m,n}^{(0)} = \frac{1}{n}\sum_{j=1}^{n} a_j^m \mathbf{X}_j, \qquad \hat{\Sigma}_{m,n}^{(0)} = \frac{1}{n}\sum_{j=1}^{n} a_j^m (\mathbf{X}_j - \hat{\mu}_{m,n})(\mathbf{X}_j - \hat{\mu}_{m,n})^T$$

as estimates for μ_m and Σ_m.

By theorem 4.3 from [10], $\hat{\mu}_{m,n}^{(0)}$, $\hat{\Sigma}_{m,n}^{(0)}$ and $\hat{\sigma}_{m,n}^{2(0)}$ are $\sqrt{n}$-consistent estimators for the corresponding parameters of the normal mixture model. This allows one to use them as pilot estimators for the one-step and EM estimators.

Now let us consider estimation of the Fisher information matrix in the case of normal mixture model. Define

$$\hat{I}_{il}(n,\tau) = \sum_{j=1}^{n} \frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau_i}\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau_l}, \tag{16}$$

$$\hat{\mathbf{I}}(n,\tau) = \left(\hat{I}_{il}(n,\tau)\right)_{i,l=1}^{R}, \qquad \hat{\mathbf{I}}(n) = \hat{\mathbf{I}}(n,\hat{\tau}) \tag{17}$$

where $\hat{\tau}$ can be any consistent estimator for τ (e.g. $\hat{\tau}_n^{OS}$ or $\hat{\tau}_n^{EM}$).

In the normal mixture model we will denote $\tau_{(l)} = (\mathbf{b}^{(l)}, \mu_l, \Sigma_l, \sigma_l^2)$, i.e. the set of all unknown parameters which describe the l-th mixture component.

Theorem 3. *Assume that the normal mixture model is taken and*

1. $\sigma_m^2 > 0$, Σ_m *are nonsingular for all* $m = 1, \dots, M$;
2. *There exist* $c > 0$ *such that for all* $j = 1, \dots, n$, $m = 1, \dots, M$, $n = 1, 2, \dots$
$$p_j^m > c.$$
3. $\tau^{(l)} \neq \tau^{(m)}$ *for all* $l \neq m$, $l, m = 1, \dots, M$.

Then

1. *There exist* $0 < c_0 < C_1 < \infty$ *such that*
$$c_o n \leq \left\|\mathbf{I}^*(n,\tau)\right\| \leq C_1 n$$
for all $n = 1, 2, \dots$.
2. $\frac{1}{n}\|\mathbf{I}^*(n,\tau) - \hat{\mathbf{I}}(n)\| \to 0$ *in probability as* $n \to \infty$.

Note. Here and below for any square matix $\mathbf{I}$ the symbol $\|\mathbf{I}\|$ means the operator norm of $\mathbf{I}$, i.e.

$$\|\mathbf{I}\| = \sup_{\mathbf{u}:\ \|\mathbf{u}\|=1} \left|\mathbf{u}^T \mathbf{I}\mathbf{u}\right|.$$

Proof. 1. At first we will show that

$$\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u} > 0 \tag{18}$$

for any τ and any $\mathbf{u} \in \mathbb{R}^P$ with $\|\mathbf{u}\| = 1$ and for any $\mathbf{p} = (p^1, \dots, p^M)$ with $p^m > c$ for all $m = 1, \dots, M$.

Recall that $\tau = (\tau_{(1)}, \dots, \tau_{(M)})$, where $\tau_{(m)}$ corresponds to the parameters describing the m-th component. Let us divide $\mathbf{u}$ into analogous blocks $\mathbf{u} = (\mathbf{u}_{(1)}^T, \dots, \mathbf{u}_{(M)}^T)^T$.

Then

$$\begin{aligned}
\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u} &= \mathsf{E}\, \mathbf{u}^T \frac{\partial}{\partial \tau} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau) \left(\frac{\partial}{\partial \tau} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)\right)^T \mathbf{u} \\
&= \mathsf{E}\left(\mathbf{u}^T \frac{\partial}{\partial \tau} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)\right)^2 \\
&= \mathsf{E}\left(\sum_{m=1}^{M} \mathbf{u}_{(m)}^T \frac{\partial}{\partial \tau_{(m)}} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)\right)^2.
\end{aligned}$$

Note that $\frac{\partial}{\partial \tau_{(m)}} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau)$ can be represented as

$$\frac{\partial}{\partial \tau_{(m)}} L(\xi_{\mathbf{p}}, \mathbf{p}, \tau) = \mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}}) \frac{p^m \varphi_{\tau_{(m)}}(\xi_{\mathbf{p}})}{f_{\mathbf{p}}(\xi_{\mathbf{p}}; \tau)} \tag{19}$$

where $\varphi_{\tau_{(m)}}$ is the normal pdf of the observation ξ from the m-th component, $f_{\mathbf{p}}$ is the pdf of the mixture defined by (10), $\mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}})$ is a vector with entries which are polynomial functions from the entries of $\xi_{\mathbf{p}}$.

Then

$$\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u} = \mathsf{E}\left(\sum_{m=1}^{M} p^m \mathbf{u}_{(m)}^T \mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}}) \varphi_{\tau_{(m)}}(\xi_{\mathbf{p}})\right)^2 \frac{1}{f_{\mathbf{p}}(\xi_{\mathbf{p}}; \tau)^2}. \tag{20}$$

Note that by the assumptions 1 and 2 of the theorem $f_{\mathbf{p}}(\xi; \tau) > 0$ for all $\xi \in \mathbb{R}^{d+1}$. Then $\mathbf{u}_{(m)}^T \mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}})$ are polynomials of $\xi_{\mathbf{p}}$ and $\varphi_{\tau_{(m)}}(\xi_{\mathbf{p}})$ are exponentials of different (due to assumption 3) and nonsingular (due to assumption 1) quadratic forms of $\xi_{\mathbf{p}}$.

Suppose that $\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u}$ for some $\mathbf{u}$ with $\|\mathbf{u}\| = 1$. Then (20) implies

$$\mathbf{u}_{(m)}^T \mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}}) = 0 \text{ a.s.} \tag{21}$$

for all $m = 1, \dots, M$.

On the other hand, (21) implies

$$\mathsf{E}\left(\mathbf{u}_{(m)}^T \mathbf{A}(\tau_{(m)}, \xi_{\mathbf{p}}) \varphi_{\tau_{(m)}}(\xi_{\mathbf{p}})\right)^2 = \mathbf{u}_{(m)}^T \mathbf{I}_{\tau_{(m)}} \mathbf{u}_{(m)} = 0,$$

where $\mathbf{I}_{\tau_{(m)}}$ is the Fisher information matrix for the unknown $\tau_{(m)}$ by one observation from the m-th component. By the assumption 1, $\mathbf{I}_{\tau_{(m)}}$ is nonsingular, so $\mathbf{u}_{(m)} = 0$ for all $m = 1, \dots, M$. This contradicts the assumption $\|\mathbf{u}\| = 1$.

So, by contradiction, (18) holds. Since $\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u}$ is a continuous function on the compact set of $\mathbf{u}$: $\|\mathbf{u}\| = 1$ and $\mathbf{p}$ satisfying assumption 2, from (18) we obtain $\mathbf{u}^T \mathbf{I}(\mathbf{p}, \tau)\mathbf{u} > c_0$ for some $c_0 > 0$. On the other hand, the representation (19) implies $\|\mathbf{I}(\mathbf{p}, \tau)\| < C_1$

Then from $\mathbf{I}^*(n, \tau) = \sum_{j=1}^n \mathbf{I}(\mathbf{p}_j, \tau)$ we obtain the first statement of the theorem.

2. To prove the second statement note that by the law of large numbers

$$\begin{aligned}\Delta_n(\tau) &= \frac{1}{n}\big(\hat{\mathbf{I}}(n, \tau) - \mathbf{I}^*(n, \tau)\big) \\ &= \frac{1}{n}\sum_{j=1}^{n}\Bigg[\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\left(\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\right)^T \\ &\quad - \mathsf{E}\,\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\left(\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\right)^T\Bigg] \\ &\xrightarrow{\mathrm{P}} 0, \qquad \text{as } n \to \infty,\end{aligned}$$

since

$$\mathsf{E}\left\|\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\right\|^4 \le C < \infty$$

for all j.

Let $B \subseteq \mathbb{R}^P$ be any open bounded neighborhood of τ. Note that

$$\mathsf{E}\sup_{\tau \in B}\left\|\frac{\partial}{\partial \tau}\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\left(\frac{\partial L(\xi_j, \mathbf{p}_j, \tau)}{\partial \tau}\right)^T\right\| < C_2 < \infty.$$

From this together with $\Delta_n(\tau) \xrightarrow{\mathrm{P}} 0$ we obtain

$$\sup_{\tau \in B}\big\|\Delta_n(\tau)\big\| \xrightarrow{\mathrm{P}} 0$$

(applying the same technique as in lemma 5.3 from [15]).

The last equation together with $\hat{\tau}_n \xrightarrow{\mathrm{P}} \tau$ implies the second statement of the Theorem. □

6 Confidence ellipsoids for $\mathbf{b}^k$

Let Ξ_n be any random dataset of size n with distribution dependent of an unknown parameter $\mathbf{b} \in \mathbb{R}^d$. Recall that a set $B_\alpha = B_\alpha(\Xi_n) \subset \mathbb{R}^d$ is called an asymptotic confidence set of the significance level α if

$$\lim_{n\to\infty} \mathsf{P}\big\{\mathbf{b} \notin B_\alpha(\Xi_n)\big\} = \alpha.$$

We will construct confidence sets for the vector of regression coefficients $\mathbf{b} = \mathbf{b}^k$ by the sample from a mixture Ξ_n described in Section 2. In the nonparametric case the set will be defined by statistics of the form

$$S^{LS}(\beta) = n\big(\beta - \hat{\mathbf{b}}^{LS}(k,n)\big)^T \hat{\mathbf{V}}_n^{-1}\big(\beta - \hat{\mathbf{b}}^{LS}(k,n)\big).$$

In the parametric case we take the matrix $\hat{\mathbf{I}}(n)$ defined by (17) and consider its inverse matrix $\hat{\mathbf{I}}(n)^{-1} = (I^-(i,m))_{i,m=1}^R$.

Note that by (16) and (17) the elements $\hat{I}_{im}$ of $\hat{\mathbf{I}}(n)$ correspond to coordinates τ_i and τ_m of the vector of unknown parameters τ. Let us take the set of indices l_m, $m = 1, \dots, d$ such that $\tau_{l_m} = b_m^k$ and consider the matrix

$$\big[\hat{\mathbf{I}}(n)^{-1}\big]_{(k)} = \big(I^-(l_i, l_m)\big)_{i,m=1}^d.$$

So, the matrix $[\hat{\mathbf{I}}(n)^{-1}]_{(k)}$ contains the elements of $\hat{\mathbf{I}}(n)^{-1}$ corresponding to $\mathbf{b}^{(k)}$ only.

Then we invert this matrix once more:

$$\hat{\mathbf{I}}_k(n)^+ = \big(\big[\hat{\mathbf{I}}(n)^{-1}\big]_{(k)}\big)^{-1}.$$

This matrix is used to construct the statistics which defines the confidence set:

$$S^{OS}(\beta) = \big(\beta - \hat{\mathbf{b}}^{OS}(k,n)\big)^T \hat{\mathbf{I}}_k(n)^+ \big(\beta - \hat{\mathbf{b}}^{OS}(k,n)\big)$$

or

$$S^{EM}(\beta) = \big(\beta - \hat{\mathbf{b}}^{EM}(k,n)\big)^T \hat{\mathbf{I}}_k(n)^+ \big(\beta - \hat{\mathbf{b}}^{EM}(k,n)\big).$$

Here $\hat{\mathbf{b}}^{OS}(k,n)$ and $\hat{\mathbf{b}}^{EM}(k,n)$ are the parts of the estimators $\hat{\tau}_n^{OS}$ and $\hat{\tau}_n^{EM}$ which esitmate $\mathbf{b}^k$.

In what follows the symbol $\star$ means any of symbols *LS*, *OS* or *EM*. The confidence set $B_\alpha^\star(\Xi_n)$ is defined by

$$B_\alpha^\star(\Xi_n) = \big\{\beta \in \mathbb{R}^d : \ S^\star(\beta) \le Q^{\chi_d^2}(1-\alpha)\big\}, \tag{22}$$

where $Q^{\chi_d^2}(1-\alpha)$ is the $(1-\alpha)$-quantile of χ^2 distribution with d degrees of freedom.

In the parametric case $\hat{\mathbf{I}}_k(n)^+$ is a positively defined matrice, so $B_\alpha^\star(\Xi_n)$ defined by (22) is the interior of an ellipsoid centered at $\hat{\mathbf{b}}^\star(k,n)$.

In the nonparametric case the matrix $\hat{\mathbf{V}}_n$ can be not positively defined for small n, so the set $B_\alpha^{LS}(\Xi_n)$ can be unbounded. We will discuss some remedial actions for this problem in Section 7.

Theorem 4. *Under the assumptions of Theorem 2,*

$$\lim_{n\to\infty} \mathsf{P}\big\{\mathbf{b}^k \notin B_\alpha^{LS}(\Xi_n)\big\} = \alpha.$$

Proof. Theorem 2 and consistency of $\hat{\mathbf{V}}_n$ imply that $S^{LS}(\mathbf{b}^k) \xrightarrow{\mathrm{W}} \chi_d^2$, so

$$\mathsf{P}\big\{\mathbf{b}^k \notin B_\alpha^{LS}(\Xi_n)\big\} = \mathsf{P}\big\{S^{LS}\big(\mathbf{b}^k\big) > Q^{\chi_d^2}(1-\alpha)\big\} \to \alpha$$

as $n \to \infty$. □

Theorem 5. *Under the assumptions of Theorem 3,*

$$\lim_{n\to\infty} \mathsf{P}\{\mathbf{b}^k \notin B_\alpha^{OS}(\Xi_n)\} = \alpha.$$

Sketch proof. By theorem 70.5 from [3] one obtains the asymptotic normality of the local MLE estimate

$$\hat{\tau}_n^{lMLE} = \operatorname{argmax}_{\tau\in D} L(\tau),$$

where D is a sufficiently small neighborhood of the true τ. Then the convergence

$$\mathbf{I}^*(n,\tau)^{-1/2}\big(\hat{\tau}_n^{OS} - \tau\big) \xrightarrow{\mathrm{W}} N(0,\mathbb{E}) \tag{23}$$

can be obtained from the asymptotic normality of $\hat{\tau}_n^{lMLE}$ by the technique of theorem 14.19 from [15].

Let us denote

$$S_0^{OS}(\beta) = \big(\beta - \hat{\mathbf{b}}^{OS}(k,n)\big)^T \mathbf{I}_k(n)^+ \big(\beta - \hat{\mathbf{b}}^{OS}(k,n)\big),$$

where $\mathbf{I}_k(n)^+$ is the theoretical counterpart of $\hat{\mathbf{I}}_k(n)^+$:

$$\mathbf{I}_k(n)^+ = \Big(\big[\mathbf{I}^*(n,\tau)^{-1}\big]_{(k)}\Big)^{-1}.$$

Then by (23), $S_0^{OS}(\mathbf{b}^k) \xrightarrow{\mathrm{W}} \chi_d^2$.

Note that (23) and the first statement of Theorem 3 imply

$$\zeta_n = \hat{\mathbf{b}}^{OS}(k,n) - \mathbf{b}^k = O_p\big(n^{-1/2}\big).$$

The second statement of Theorem 3 implies

$$\frac{1}{n}\big\|\hat{\mathbf{I}}_k(n)^+ - \mathbf{I}_k(n)^+\big\| \xrightarrow{\mathrm{P}} 0.$$

So

$$S_0^{OS}\big(\mathbf{b}^k\big) - S^{OS}\big(\mathbf{b}^k\big) = \zeta_n^T \frac{1}{n}\big(\hat{\mathbf{I}}_k(n)^+ - \mathbf{I}_k(n)^+\big)\zeta_n \xrightarrow{\mathrm{P}} 0$$

and $S^{OS}(\mathbf{b}^k) \xrightarrow{\mathrm{W}} \chi_d^2$.

This completes the proof. □

7 Results of simulations

We carried out a small simulation study to assess performance of the parametric and nonperametric confidence intervals described above. A two component mixture ($M = 2$) of simple regressions was simulated. The regression models were of the form

$$Y = b_0^\kappa + b_1^\kappa X + \varepsilon^\kappa, \tag{24}$$

where $X^k \sim N(\mu_k, \Sigma_k^2)$ and Y are the observed regressor and response, κ is the unobserved number of components, ε^k is the regression error. The error ε^k has zero mean and variance σ_k^2.

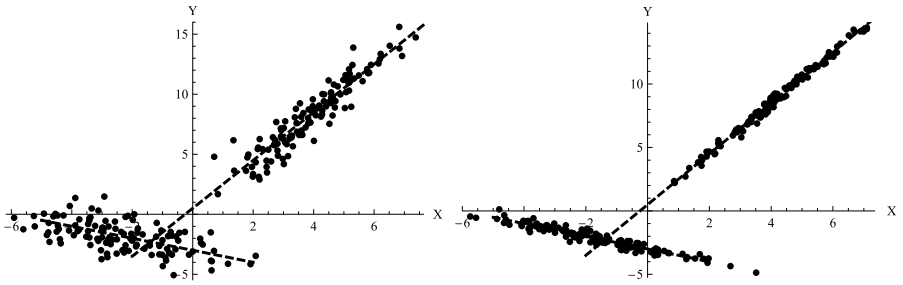

Fig. 1. Typical scatterplots of data in Experiment 1 (left) and Experiment 2 (right)

Table 1. Parameters for simulation in Experiments 1 and 2

k	1	2
μ_k	-2	4
Σ_k	3	2
σ_k	1	1
b_0^k	-3	0.5
b_1^k	-0.5	2

The mixing probabilities were simulated by the following stochastic model:

$$p_{j;N}^m = \frac{u_j^m}{\Sigma_{s=1}^M u_j^s},$$

where u_j^m are independent uniformly distributed on [0, 1].

For each sample size n we generated 1000 samples. Parametric (EM) and nonparametric (LS) confidence ellipsoids were constructed by each sample. The parametric ellipsoids were based on EM-estimates which used the LS-estimates as the pilot ones and $\hat{\mathbf{I}}_k(n)^+$ as the matrix for the quadratic form in S^{EM}.

The nonparametric confidence ellipsoids were based on the LS-estimates. As it was mentioned in Section 6, the matrix $\hat{\mathbf{V}}_n$ can be not positively defined. Then the corresponding confidence set will be unbounded. In the case of simple regression (24) this drawback can be cured by the use of improved weights b_j^+ defined in [8] instead of a_j^k in (12)–(14). This technique was used in our simulation study.

All the ellipsoids were constructed with the nominal confidence level $\alpha = 0.05$. The frequencies of covering true $\mathbf{b}^k$ by the constructed ellipsoids and their mean volume were calculated in each simulation experiment.

Experiment 1. The values of parameters for this experiment are presented in Table 1. The errors ε^k were Gaussian. This is a "totally separated" model in which the observations can be visually divided into two groups corresponding to different mixture components (see the left panel at Fig. 1).

Covering frequencies and mean volumes of the ellipsoids for different sample sizes n are presented in Table 2. They demonstrate sufficient accordance with the nominal significance level for sample sizes greater then 1000. Extremely large mean

Table 2. Experiment 1 results (k is the number of component)

Covering frequencies

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	0.954	0.821	0.946	0.951
10^3	0.975	0.914	0.947	0.952
10^4	0.988	0.951	0.95	0.95
10^5	0.951	0.963	0.952	0.953
10^6	0.936	0.949	0.936	0.951

Average volume of ellipsoids

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	$298 * 10^6$	$262 * 10^6$	2.177543	0.243553
10^3	1364	394	0.186303	0.021234
10^4	0.476327	0.317320	0.018314	0.002062
10^5	0.041646	0.030047	0.001845	0.000207
10^6	0.004121	0.002988	0.000185	0.000021

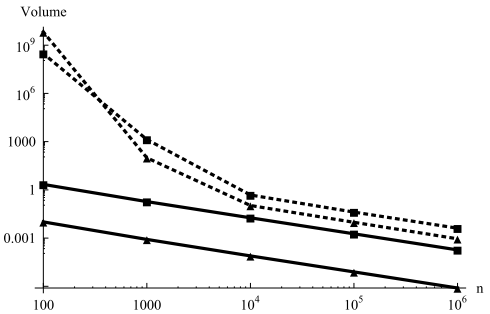

Fig. 2. Average volumes of ellipsoids in Experiment 1 (■) and Experiment 2 (▲). Solid lines for EM, dashed lines for LS (First component)

volumes for the LS-ellipsoids are due to poor performance of the estimates $\hat{\mathbf{V}}_n$ for small and moderate sample sizes n.

The parametric confidence sets are significantly smaller then the nonparametric ones.

Experiment 2. To see how the standard deviations of regression errors affect the performance of our algorithms we reduced them to $\sigma_k = 0.25$ in the second experiment, keeping all other parameters unchanged. A typical scatterplot of such data is presented on the right panel of Fig. 1.

The results of this experiment are presented in Table 3. They are compared graphically to the results of Experiment 1 in Fig. 2. The covering frequencies are not significantly changed. In comparison to Experiment 1, the average volumes decreased significantly for EM-ellipsoids but not for the LS ones.

Experiment 3. Here we consider another set of parameters (see Table 4). The regression errors are Gaussian. In this model the subjects cannot be classified uniquely by their observed variables (see the left panel in Fig. 3).

Table 3. Experiment 2 results (k is the number of component)

Covering frequencies

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	0.886	0.922	0.950	0.943
10^3	0.942	0.910	0.945	0.948
10^4	0.954	0.946	0.951	0.955
10^5	0.962	0.958	0.943	0.950
10^6	0.955	0.937	0.961	0.942

Average volume of ellipsoids

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	6560085466	43879747920	0.01022148	0.01199164
10^3	98.92863	182799.39295	0.0008286214	0.0011644647
10^4	0.1061491	0.2465760	7.846928×10^{-05}	1.196875×10^{-04}
10^5	0.009584066	0.021603033	7.870776×10^{-06}	1.191609×10^{-05}
10^6	0.0009045894	0.0021206426	7.875141×10^{-07}	1.189581×10^{-06}

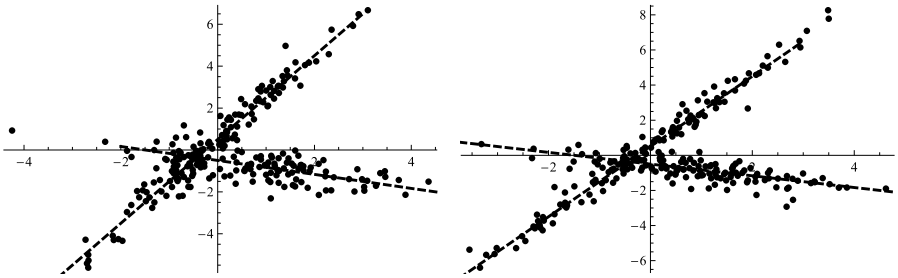

Fig. 3. Typical scatterplots of data in Experiment 3 (left) and Experiment 4 (right)

The results are presented in Table 5. Again, the EM-ellipsoids outperform the LS ones.

Experiment 4. In this experiment the parameters are the same as in Experiment 3, but the regression errors are **not** Gaussian. We let $\varepsilon^k = \sqrt{3/5}\sigma_k\eta$, where η has the Student-T distribution with 5 degrees of freedom. So the errors here have the same variances as in Experiment 3, but their distributions are heavy-tailed. Note that 5 is the minimal number of degrees of freedom for which the assumption $\mathsf{E}(\varepsilon_k)^4$ of Theorem 2 holds.

A typical data scatterplot for this model is presented on the right panel of Fig. 3. It is visually indistinguishable from the typical pattern of the Gaussian model from Experiment 3, presented on the left panel.

Results of this experiment are presented in Table 6. Note that in this case the covering proportion of the EM-ellipsoids does not tend to the nominal $1 - \alpha = 0.95$ for large n. The covering proportion of LS-ellipsoids is much nearer to 0.95. So the heavy tails of distributions of the regression errors deteriorate performance of (Gaussian model based) EM-ellipsoids but not of nonparametric LS-ellipsoids.

Table 4. Parameters for simulation in Experiments 3 and 4

k	1	2
μ_k	0	1
Σ_k	2	2
σ_k	0.5	0.5
b_0^k	0.5	−0.5
b_1^k	2	$-\frac{1}{3}$

Table 5. Experiment 3 results (k is the number of component)

Covering frequencies

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	0.920	0.928	0.949	0.935
10^3	0.953	0.943	0.948	0.946
10^4	0.951	0.957	0.954	0.945
10^5	0.947	0.963	0.942	0.961
10^6	0.945	0.951	0.948	0.939

Average volume of ellipsoids

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	294.5016	28837.3340	0.05897494	0.06088698
10^3	0.6088472	0.6274452	0.005250821	0.004937218
10^4	0.05837274	0.05594969	0.0005024635	0.0004993278
10^5	0.005604424	0.00551257	4.987135×10^{-05}	5.024126×10^{-05}
10^6	0.0005625693	0.0005550716	4.978973×10^{-06}	5.029275×10^{-06}

8 An application to sociological data analysis

To demonstrate possibilities of the developed technique, we present a toy example of construction of confidence ellipsoids in statistical analysis of dependence between school performance of students and political attitudes of their adult environment. The analysis was based on two data sets. The first one contains results of the External independent testing in Ukraine in 2016 – EIT-2016. EIT is a a set of exams for high schools graduates for admission to universities. Data on EIT-2016[2] contain individual scores of examinees with some additional information including the region of Ukraine at which the examinee's school was located. The scores range from 100 to 200 points.

We considered the information on the scores on two subjects: *Ukrainian language and literature* (Ukr) and on *Mathematics* (Math). EIT-2016 contains data on these scores for nearly 246 000 examinees. It is obvious that Ukr and Math scores should be dependent and the simplest way to model this dependency is the linear regression:

$$\mathrm{Ukr} = b_0 + b_1 \mathrm{Mat} + \varepsilon.$$

We suppose that the coefficients b_0 and b_1 may depend on the political attitudes of the adult environment in which the student was brought up. Say, in a family of Ukrainian

[2]Taken from the official site of *Ukrainian Center for Educational Quality Assessment* https://zno.testportal.com.ua/stat/2016.

Table 6. Experiment 4 results (k is the number of component)

Covering frequencies

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	0.912	0.915	0.943	0.937
10^3	0.948	0.945	0.949	0.959
10^4	0.937	0.945	0.929	0.953
10^5	0.947	0.951	0.915	0.930
10^6	0.961	0.953	0.634	0.763

Average volume of ellipsoids

n	LS		EM	
	$k=1$	$k=2$	$k=1$	$k=2$
100	997.1288	584.8507	0.06740671	0.06419959
10^3	0.7006367	0.6127510	0.005262779	0.004971307
10^4	0.05798962	0.05624429	0.0004850319	0.0004884329
10^5	0.005621060	0.005574176	4.667732×10^{-05}	4.746252×10^{-05}
10^6	0.0005616700	0.0005566846	4.666926×10^{-06}	4.745760×10^{-06}

independence adherents one expects more interest to Ukrainian language than in an environment critical toward the Ukrainian state existence.

Of course EIT-2016 does not contain any information on political issues. So we used the second data set with the official data on the results[3] of the Ukrainian Parliament elections-2014 to get approximate proportions of adherents of different political choices in different regions of Ukraine.

29 political parties and blocks took part in the elections. The voters were able also to vote against all or not to take part in the voting. We divided all the population of voters into three components:

(1) Persons who voted for parties which then created the ruling coalition (BPP, People's front, Fatherland, Radical party, Self Reliance). This is the component of persons with positive attitudes to the pro-European Ukrainian state.

(2) Persons who voted for the Opposition block, voters against all, and voters for small parties which where under 5% threshold at these elections. These are voters critical to the pro-European line of Ukraine but taking part in the political life of the state.

(3) Persons who did not take part in the voting. These are persons who did not consider Ukrainian state as their own one or are not interested in politics at all.

We used the results of elections to calculate the proportions of each component in each region of Ukraine where the voting was held. These proportions were taken as estimates for the probabilities that a student from a corresponding region was brought up in the environment of a corresponding component. That is, they were considered as the mixing probabilities.

The LS- and EM-ellipsoids for b_0 and b_1 obtained by these data are presented on Fig. 4. The ellipsoids were constructed with the significance level $\alpha = 0.05/3 \approx 0.01667$, so by the Bonferroni rule, they are unilateral confidence sets with $\alpha = 0.05$.

[3] See the site of *Central Election Commission (Ukraine)* http://www.cvk.gov.ua/vnd_2014/.

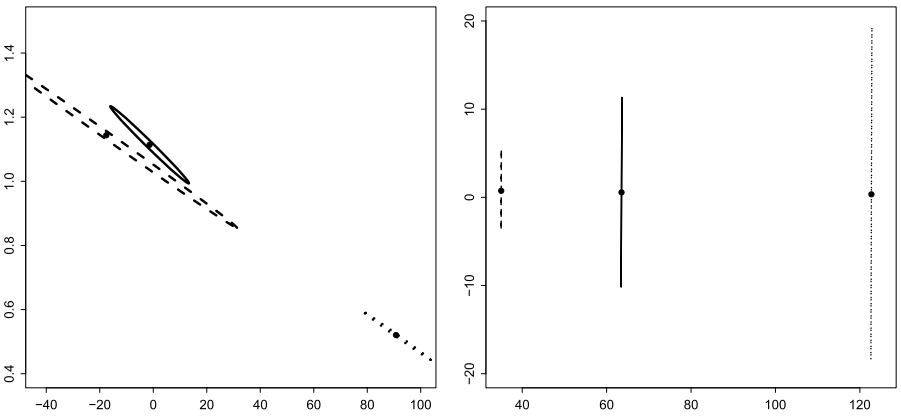

Fig. 4. LS (left) and EM (right) confidence ellipsoids for the EIT data. Components: (1) dotted line, (2) dashed line, (3) solid line. b_0 on the horizontal axis, b_1 on the vertical axis

Since the ellipsoids are not intersecting in both cases, one concludes that the vectors of regression coefficients $(b_0^i, b_1^i), i = 1, \ldots, 3$ are significantly different for different components.

Note that the EM approach leads to estimates significantly different from the LS ones. This may suggest that the normal mixture model (3) does not hold for the data. Does the nonparametric model hold for them? Analysis of this problem and meaningful sociological interpretation of these results lie beyond the scope of this article.

9 Concluding remarks

We considered two approaches to the construction of confidence sets for coefficients of the linear regression in the mixture model with varying mixing probabilities. Both approaches demonstrate sufficient agreement of nominal and real significance levels for sufficiently large samples when the data satisfy underlying assumptions of the confidence set construction technique. The parametric approach needs a significant additional a priori information in comparison with the nonparametric one. But it utilizes this information providing much smaller confidence sets than in the nonparametric case.

On the other hand, the nonparametric estimators proved to be a good initial approximation for the construction of parametric estimators via the EM-algorithm. Nonparametric confidence sets also perform adequately in the cases when the assumptions of parametric model are broken.

Acknowledgments

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

The research was supported in part by the Taras Shevchenko National University of Kyiv scientific grant N 16БФ038-02.

References

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

[2] Benaglia, T., Chauveau, D., Hunter, D.R., mixtools, Y.D.S.: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software **32**(6), 1–29 (2009). https://doi.org/10.18637/jss.v032.i06

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

[4] Faria, S.: Soromenhob Gilda Fitting mixtures of linear regressions. Journal of Statistical Computation and Simulation **80**(2), 201–225 (2010). MR2757044. https://doi.org/10.1080/00949650802590261

[5] Grün, B., Friedrich, L.: 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). MR2173118

[6] Liubashenko, D., Maiboroda, R.: Linear regression by observations from mixture with varying concentrations. Modern Stochastics: Theory and Applications **2**(4), 343–353 (2015). MR3456142. https://doi.org/10.15559/15-VMSTA41

[7] Maiboroda, R.: Statistical analysis of mixtures. Kyiv University Publishers, Kyiv (2003). (in Ukrainian)

[8] Maiboroda, R., Kubaichuk, O.: Asymptotic normality of improved weighted empirical distribution functions. Theor. Probab. Math. Stat. **69**, 95–102 (2004). MR2110908. https://doi.org/10.1090/S0094-9000-05-00617-4

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

[10] Maiboroda, R., Sugakova, O.: Statistics of mixtures with varying concentrations with application to DNA microarray data analysis. Journal of nonparametric statistics. **24**(1), 201–205 (2012). MR2885834. https://doi.org/10.1080/10485252.2011.630076

[11] Maiboroda, R.E., Sugakova, O.V., Doronin, A.V.: Generalized estimating equations for mixtures with varying concentrations. The Canadian Journal of Statistics **41**(2), 217–236 (2013). MR3061876. https://doi.org/10.1002/cjs.11170

[12] Maiboroda, R., Sugakova, O.: Sampling bias correction in the model of mixtures with varying concentrations. Methodology and Computing in Applied Probability. **17**(1) (2015). MR3306681. https://doi.org/10.1007/s11009-013-9349-4

[13] McLachlan, G.: Krishnan Thriyambakam The EM Algorithm and Extensions, 2nd edn. Wiley, (2008). MR2392878. https://doi.org/10.1002/9780470191613

[14] Seber, G.A.F., Lee, A.J.: Linear Regression Analysys. Wiley, (2003). MR1958247. https://doi.org/10.1002/9780471722199

[15] Shao, J.: Mathematical statistics. Springer, New York (1998). MR1670883

[16] Titterington, D.M., Smith, A.F., Makov, U.E.: Analysis of Finite Mixture Distributions. Wiley, New York (1985). MR0838090