A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 3(314)2015
http://dx.doi.org/10.18778/0208-6018.314.03

Oleksii Doronin[*], *Rostislav Maiboroda*[**]

GEE ESTIMATORS IN MIXTURE MODEL WITH VARYING CONCENTRATIONS

Abstract. We discuss a semiparametric mixture model where some components are parameterized with common Euclidean parameter and others are fully unknown. We introduce GEE (generalized estimating equations) approach and adaptive GEE-based approach for parameter estimation. Derived estimators are consistent and asymptotically normal, and they are optimized in terms of their dispersion matrices. Proposed techniques are tested on simulated samples.

Keywords: mixture model, semiparametric estimation, GEE.

1. INTRODUCTION

The cumulative distribution function (CDF) of one observation in a mixture model is expressed by a linear combination of some CDFs $F_1,...,F_M$ with probabilities $p^1,...,p^M$, $\sum_{m=1}^{M} p^m = 1$ (i.e. $F_\xi(x) = \sum_{m=1}^{M} p^m F_m(x)$). Note that F_m is called the CDF of the m-th mixture component, and p^m – the component concentration. In mixture model with varying concentrations p^m depends on the observation index: $p^m = p_j^m$, $j = \overline{1,N}$. Thus,

$$F_{\xi_j}(x) = \sum_{j=1}^{M} p_j^m F_m(x), \quad j = \overline{1,N}.$$

We consider the case when some parametric model is known for the first K components: $F_m(x) = F_m(x;t)$, $m = \overline{1,K}$. Parameter t is assumed to be Euclidean: $t \in \Theta \subset \Re^d$. The true value of t we designate as ϑ and assume that it is unknown. The CDFs of the last $M-K$ mixture components are assumed to

[*] Ph.D. student, Department of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, Taras Shevchenko National University of Kyiv.

[**] Ph.D., Department of Probability Theory, Statistics and Actuarial Mathematics, Mechanics and Mathematics Faculty, Taras Shevchenko National University of Kyiv.

be fully unknown. We also assume that concentrations p_j^m are known. Our goal is to estimate ϑ. To do this, we derive consistent and asymptotically normal estimators, and optimize them in terms of their dispersion matrices.

2. NONPARAMETRIC ESTIMATE FOR DISTRIBUTION FUNCTION

CDF of the *m*-th component may be estimated through the weighted empirical distribution function:

$$\hat{F}_m(x) := \frac{1}{N}\sum_{j=1}^{N} a_j^m I_{\{\xi_j \le x\}}.$$

Weights a_j^m are taken as the solution of the minimization problem of maximal variance of unbiased estimates of $F_m(x)$ for all possible CDFs F_m (i.e. $a^m = p\Gamma^{-1}e_m$ where $p := (p_j^m)_{j=\overline{1,N},m=\overline{1,M}} \in \Re^{N\times M}$, $\Gamma := \frac{1}{N}p^T p \in \Re^{M\times M}$, $e_m := (\chi\{i=m\})_{i=1,\dots,M}$). See Maiboroda et al. (2008) for details.

Note that weights a_j^m can be negative. Thus, we can improve $\hat{F}_m(x)$ by introducing improved empirical distribution function (see Maiboroda et al. (2005)):

$$\hat{F}_m^+(x) := \min(1, \max_{y\le x} F(y)).$$

3. GEE ESTIMATE

Consider some set of measurable functions $g_1(\xi;t),\dots,g_K(\xi;t) \to \Re^d$. Theoretical moment $\int g_k(x;t)F_k(dx)$ may be estimated by the weighted empirical moment as

$$\hat{g}_k^k(t) := \frac{1}{N}\sum_{j=1}^{N} a_j^k g_k(\xi_j;t).$$

Define the joint weighted empirical moment of $\hat{g}_k^k(t)$ as

$$\hat{g}(t) := \sum_{k=1}^{K} \hat{g}_k^k(t).$$

Definition. GEE estimator $\hat{\vartheta}$ for ϑ is the measurable function from sample $\xi_1,...,\xi_N$ such that $\hat{g}(\hat{\vartheta}) = 0$. Next we assume that $P[\exists t \in \Theta : \hat{g}(\hat{\vartheta}) = 0] \to 1$ as $N \to \infty$.

Example. Moment estimators can be represented as GEE estimators. Let $h_1,...,h_K$ be the set of estimating functions. Denote theoretical moment of $h_k(x)$ as $H_k(t) := \int h_k(x) F_k(dx;t)$, $k = \overline{1,K}$. Define estimating functions as $g_k(x;t) := h_k(x) - H_k(t)$, $k = \overline{1,K}$. GEE estimator $\hat{\theta}$ can be represented as $\hat{\vartheta} := H^{-1}(\sum_{k=1}^{K} \hat{h}_k^k)$ where H^{-1} is the inversed function to $H(t) := \sum_{k=1}^{K} H_k(t)$. Analogous improved moment estimate with $\hat{h}_k^k := \int h_k(x) \hat{F}_k^+(dx)$ can be introduced.

Consistency for moment estimators is shown in theorem 3.1 from Doronin (2014a).

4. ASYMPTOTICS OF GEE ESTIMATOR

Assume that CDFs $F_1,...,F_M$ are absolutely continuous with respect to sigma-finite measure μ on the space of observations. Denote densities of each component's distributions as $f_k(x) := \frac{dF_k(x;\vartheta)}{d\mu(x)}$, $k = \overline{1,K}$, $f_k(x) := \frac{dF_k(x)}{d\mu(x)}$, $k = \overline{K+1,M}$.

Introduce the matrix of estimating functions

$$G(x) := \begin{pmatrix} g_1(x;\vartheta) \\ \vdots \\ g_K(x;\vartheta) \end{pmatrix} \in \Re^{K \times d}.$$

Expectation of $G(x)$ from the *m*-th component designate as $\overline{G_m} := \int G(x) F_m(dx)$, $m = 1,...,M$.

Introduce the following notations.

$$\alpha_{r,s} := (\alpha_{r,s}^{k,l})_{k,l=\overline{1,K}} := \left(\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} a_j^k a_j^l p_j^r p_j^s \right)_{k,l=\overline{1,K}} \in \Re^{K \times K},\ r,s = \overline{1,M}.$$

$$\beta_m := (\beta_m^{k,l})_{k,l=\overline{1,K}} := \left(\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} a_j^k a_j^l p_j^m \right)_{k,l=\overline{1,K}} \in \Re^{K\times K},\ m = \overline{1,M}.$$

$$R(x) := \sum_{m=1}^{M} \beta_m f_m(x) \in \Re^{K\times K}.$$

$$Z := \int G(x)^T R(x) G(x) \mu(dx) - \sum_{r,s=1}^{M} \overline{G}_r^T \alpha_{r,s} \overline{G}_s \in \Re^{d\times d}.$$

$$V := \sum_{k=1}^{K} \int \frac{\partial g_k(x;t)}{\partial t}\Big|_{t=\vartheta} F_k(dx) \in \Re^{d\times d}.$$

Theorem 4.1. (Theorem 3.4 from Doronin (2014a)) Let $\hat{\vartheta}$ be GEE estimator in introduced definitions, and U be some open neighborhood of the true parameter value ϑ. Assume the following.:

(i) $\hat{\vartheta}$ converges in probability to ϑ as $N \to \infty$.

(ii) Derivatives $g_k'(x;t) = \partial g_k(x;t)/\partial t^T$ exist and are integrable (i.e. $E_t[\| g_k'(\eta_m;t) \|] < \infty$) for $t \in U$, where E_t denotes expectation under condition that the true parameter value is t, and η_m are the formal random values with distrubutions F_m.

(iii) Functions $\overline{g}_k^m(t) = E_\vartheta[g_k(\eta_m;t)]$ are continuous on U.

(iv) $E_\vartheta[\sup_{t\in U} \| g_k(\eta_m;t) \|] < \infty$.

(v) Limit matrix Γ exist and is nonsingular.

(vi) Matrices $\alpha_{r,s}$ and β_m exist.

(vii) Matrix V is nonsingular.

(viii) GEE is unbiased, i.e. $\sum_{k=1}^{K} E_t[g_k(\eta_k;t)] = 0$ for $t \in U$.

Then $\sqrt{N}(\hat{\vartheta} - \vartheta)$ converges in distribution to Gaussian distribution with zero mean and covariance matrix $V^{-1} Z V^{-T}$.

5. LOWER BOUND OF DISPERSION MATRIX FOR GEE ESTIMATOR

Assume that the matrix Z and nonsingular matrix V exist. Without loss of generality we can assume that two conditions for GEE estimator $\hat{\vartheta}$ are fulfilled:

(i1) $\int g_k(x;\vartheta) F_k(dx) = 0$, $k = \overline{1,K}$ (unbiasedness);

(i2) matrix V is the unit matrix.

Consider the minimization problem of dispersion matrix Z in Loewner ordering (i.e. $A \geq B$ if $A - B$ is non-negatively defined) over all $g_k(x;\vartheta)$ satisfying conditions **(i1)**, **(i2)**. Thus, we have to minimize $c^T Z c$ for all $c \in \Re^d$. The solution of this problem is the set of estimating functions $g_k^*(x;\vartheta)$, which give us the lower bound of dispersion matrix Z^* (see theorem 4.1 from Doronin (2014a)).

6. ADAPTIVE ESTIMATE

Unfortunately, it is impossible to use in practice the optimal estimating functions $g_k^*(x;\vartheta)$, which give the lower bound of dispersion matrix. The first reason is that they depend on unknown densities $f_k(x)$, $k = \overline{1,K}$. The second one is the difficulty to solve the GEE in the general case. Therefore, we consider the adaptive approach.

Each function $g_k(x;t)$ can be approximated as $B_k u_k(x;t)$ where $B_k \in \Re^{d \times L_k}$ is some matrix of coefficients to be found, and $u_k(x;t) \in \Re^{L_k}$ is the vector of some predefined basis functions (e.g. B-splines). Under conditions **(i1)**, **(i2)** equation $\sum_{k=1}^{K} \hat{g}_k^k(t) = 0$ we can approximate as

$$0 = \sum_{k=1}^{K} \hat{g}_k^k(t) \approx \sum_{k=1}^{K} B_k \hat{u}_k^k(\vartheta) + (t - \vartheta).$$

The solution of this approximated equation is $t = \vartheta - \sum_{k=1}^{K} B_k \hat{u}_k^k(\vartheta)$. Thus, one can start with some consistent estimate $\widetilde{\vartheta}$ and define adaptive estimate as

$$\hat{\vartheta} := \widetilde{\vartheta} - \sum_{k=1}^{K} B_k \hat{u}_k^k(\widetilde{\vartheta}).$$

Consistency and asymptotic normality of introduced adaptive estimate is shown in lemma 3.3 from Doronin (2014b).

7. NUMERICAL RESULTS

We chose a three-component mixture model to simulate. All components are taken Gaussian, with parameter values (m,σ) as (−3.2), (3.2), (0.2), for each component, respectively. The first two components are assumed to be

parameterized with $\vartheta = (m_1, m_2, \sigma)^T$ (different means, common standard deviation). Distribution of the third component is assumed to be fully unknown. Concentrations were also generated as the pseudo-random values, derived by formula $p_j^m = s_j^m / (s_j^1 + s_j^2 + s_j^3)$ where s_j^m is taken from uniform distribution on $[0,1]$. Series of samples with sizes 50, 100, 250, 500, 750, 1000, 2000, 5000 were simulated, 2000 samples in each series. Vectors of basis functions $u_k(x;t)$ for adaptive estimate were chosen as the set of uniform cubic B-splines with knots at points $m + i\sigma$, where m and σ are the mean and standard deviation of the k-th component, respectively, $i = -5,...,5$. Matrices B_k were chosen to minimize dispersion matrix. Results are shown in Figure 1.

CONCLUSIONS

The mixture model with varying concentrations is considered. Several estimators for this model are introduced (moment, GEE, adaptive). The proposed estimators are consistent and asymptotically normal under some conditions. Performance of moment and adaptive estimators are compared on simulated samples. Dispersion of introduced estimators converges to its theoretical asymptotic value for samples with 1000 and more observations.

REFERENCES

Doronin O. (2014a), *Lower bound of dispersion matrix for semiparametric estimation in mixture model*. "Theory of Probability and Mathematical Statistics", no. 90, p. 64−76.

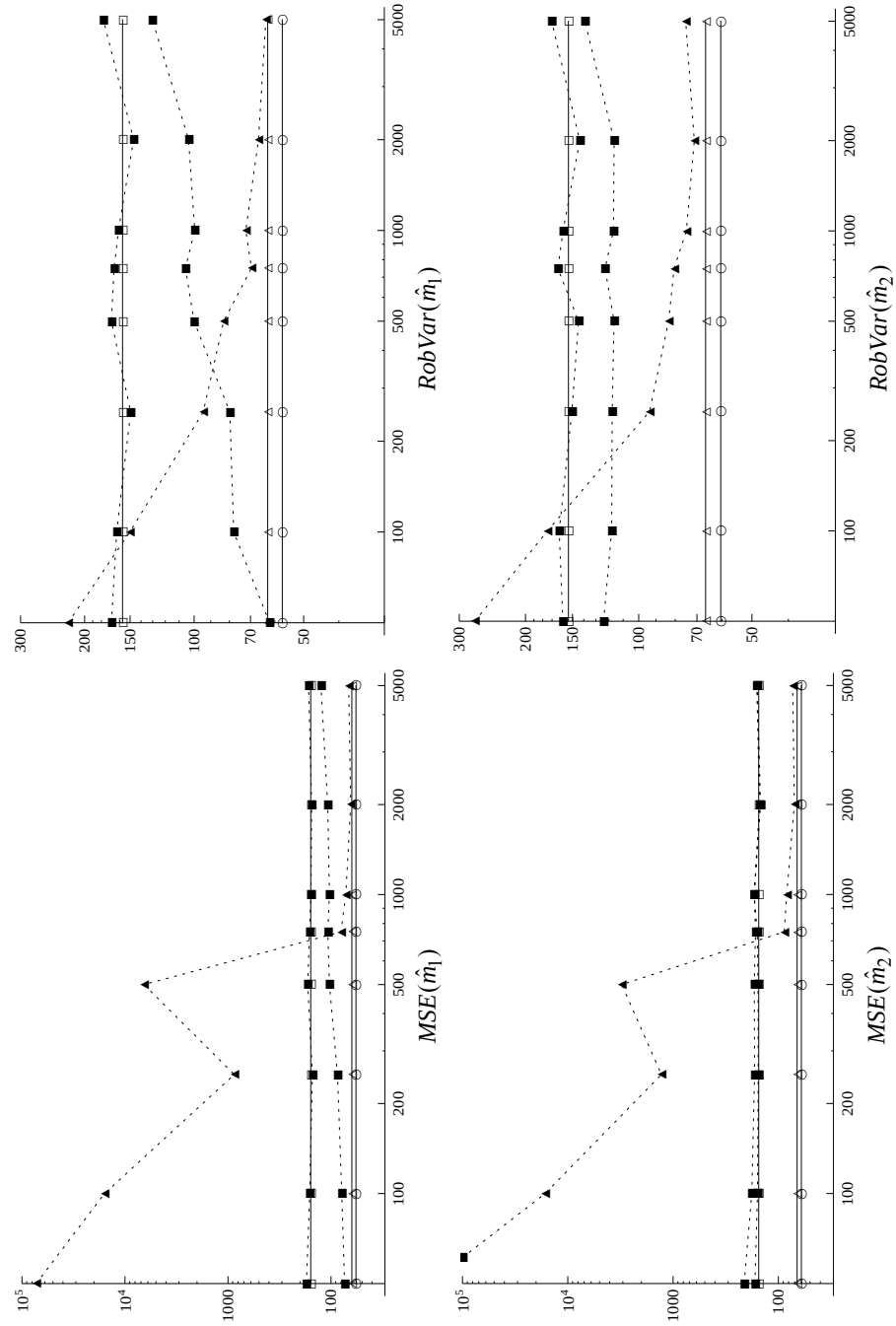
Doronin O. (2014b), *Adaptive estimation in semiparametric model of mixture with varying concentrations*. "Theory of Probability and Mathematical Statistics", no. 91, p. 27−38.

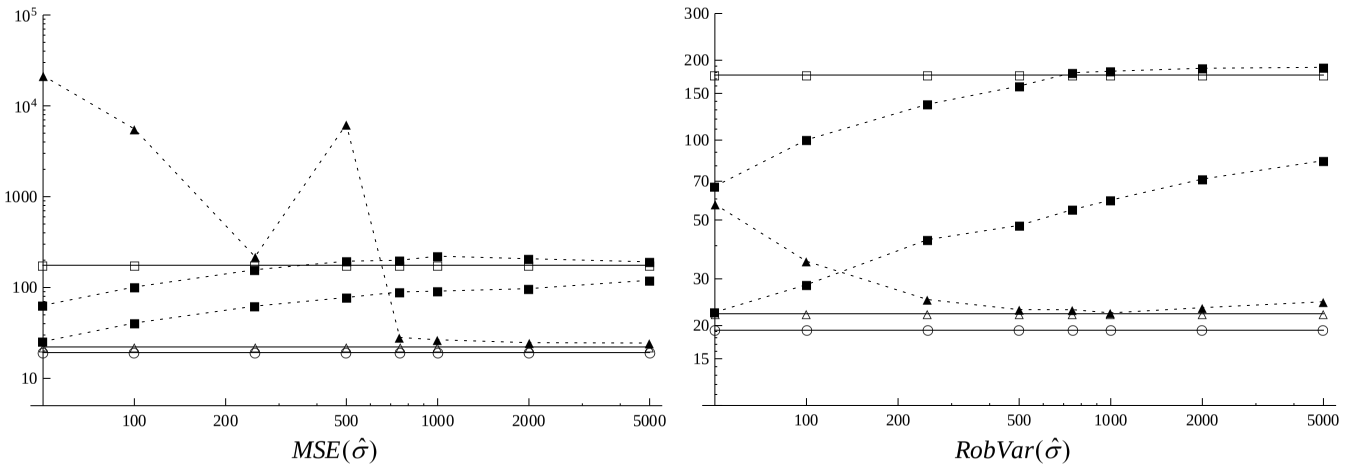
Doronin O. (2012), *Robust Estimates for Mixtures with Gaussian Component*. "Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics & Mathematics" (in Ukrainian), vol. 1, p. 18–23.

Maiboroda R., Sugakova O. (2008), *Estimation and classification by observations from mixtures.* Kyiv University Publishers, Kyiv (in Ukrainian).

Maiboroda R., Kubaichuk O. (2005), *Improved estimators for moments constructed from observations of a mixture.* "Theory of Probability and Mathematical Statistics", no. 70, p. 83−92.

Maiboroda R., Sugakova O., Doronin A. (2013), *Generalized estimating equations for mixtures with varying concentrations.* "The Canadian Journal of Statistics", no. 41, vol. 2, p. 217−236.


$RobVar(\hat{m}_1)$
$RobVar(\hat{m}_2)$
$MSE(\hat{m}_1)$
$MSE(\hat{m}_2)$



Here MSE is the mean squared error of the parameter estimate multiplied by number of observations N. $RobVar$ is the robust estimate of MSE through the interquartile range of parameter estimate. Symbol ■ indicates the moment estimates (lower line for improved and upper line for unimproved), and ▲ − adaptive estimates. White symbols indicate theoretical dispersion. Symbol ○ indicates the lower bound.

Figure 1. Dispersion of estimates

Source: plots are generated by Wolfram Mathematica using our own script.