# SEMIPARAMETRIC ESTIMATION IN MIXTURE MODELS WITH VARYING MIXING PROBABILITIES

A.V. Doronin
*Kyiv National Taras Shevchenko University*
*Kyiv, Ukraine*
e-mail: al_doronin@ukr.net

**Abstract**

We consider a model of mixture with concentrations varying from observation to observation. Semiparemetric estimation problems are considered for this model. We present three types of estimators, namely the moment, quantile and adaptive ones. Their performance is compared both analitically and by simulations.

## 1 Introduction

We consider the series of samples from $N$ observations $\{\xi_{1;N}, ..., \xi_{N;N}\}_{N\geq 1}$. Each observation $\xi_{j;N}$ can have a CDF $F_m(\cdot)$, $m = \overline{1,M}$ with some predefined probability (concentration) $p_{j;N}$. Thus, the CDF of $\xi_{j;N}$ is $P[\xi_{j;N} \in A] = \sum_{m=1}^{M} p_{j;N}^m F_m(A)$, $A \subset \mathbb{R}$.

In what follows we assume that the CDF of the first component is parametrized with some Euclidean parameter $t \in \Theta \subset \mathbb{R}^d$ (i.e. $F_1(A) = F_1(A,t)$). The true value of parameter we denote as $\vartheta \in \Theta$. The CDFs of the rest of the components are assumed to be fully unknown.

Moment, quantile and adaptive estimators for $\vartheta$ are discussed in Sections 2-4. Performance of these estimators is assessed via simulations in Section 5.

## 2 Moment estimators

Weighted empirical CDF $\hat{F}_{m;N}(x)$ with minimax weight coefficients $a_{j;N}^m$ is defined in [4] as $\hat{F}_{m;N}(x) := \frac{1}{N}\sum_{j=1}^{N} a_{j;N}^m \mathbb{I}_{\{\xi_{j;N}\leq x\}}$.

Improved weighted empirical CDF is introduced in [3] as $\hat{F}^+_{m;N}(x) := \min\left\{1, \sup_{y\leq x} \hat{F}_{m;N}(y)\right\}$.

Consider some measurable function $h : \mathbb{R} \to \mathbb{R}^d$.

We define moment estimator as a solution of moment equation

$$\hat{\vartheta}_N^{simple} := \arg\limits_{t\in\Theta} \left\{ \int h(x)\hat{F}_{1;N}(dx) = \int h(x)F_1(dx,t) \right\}. \tag{1}$$

Alternatively, we define improved moment estimator

$$\hat{\vartheta}_N^{impr} := \arg\limits_{t\in\Theta} \left\{ \int h(x)\hat{F}^+_{1;N}(dx) = \int h(x)F_1(dx,t) \right\}. \tag{2}$$

Consistency and asymptotic normality are demonstrated for both $\hat{\vartheta}_N^{simple}$ and $\hat{\vartheta}_N^{impr}$ in [4] and [3].

# 3 Quantile estimators

**Estimators for quantiles**

It is proposed in [4] to define an estimator for a quantile $\hat{Q}_{m;N}(\alpha)$ of level $\alpha$ as a value of a function, inversed to piece-wise linear interpolation of improved CDF $\hat{F}^{+}_{m;N}$ defined in section 2. Consistency and asymptotic normality of this estimate are demonstrated in [4].

Asymptotic normality of a vector $(\hat{Q}^m_N(\alpha_i))_{i=\overline{1,q}}$ is shown in [2], and it is found its dispersion matrix.

**Theorem 1.** *(Theorem 1 from [2])*

*Assume that*

*1.* $\sup\limits_{j;N} |a^m_{j;N}| < \infty$.

*2. The limits* $\langle p^k p^l (a^m)^2 \rangle$, $\langle p^k (a^m)^2 \rangle$ *exist for all* $k, l = 1, M$.

*3.* $F_k(\cdot)$ *are continuous on* $\mathbb{R}$, $k = 1, M$.

*4. The unbiasedness condition* $\langle a^m p^k \rangle = \mathbb{I}_{k=m}$, $k = 1, M$ *holds.*

*5. Functions* $F_k(\cdot)$, $k = 1, M$ *are monotone increasing in some neighborhoods* $I_1, ..., I_q$ *of points* $Q^{F_m}(\alpha_1), ..., Q^{F_m}(\alpha_q)$ *respectively.*

*6. Function* $F_m(\cdot)$ *has a continuous derivative* $f_m(\cdot)$ *on* $I_1, ..., I_q$, *and* $f_m(Q^{F_m}(\alpha_i)) \neq 0$, $i = 1, q$.

*Then the vector* $(\sqrt{N}(\hat{Q}^m_N(\alpha_i) - Q^{F_m}(\alpha_i)))_{i=\overline{1,q}}$ *weakly converges as* $N \to \infty$ *to a Gaussian random vector with zero mean and covariance matrix* $S = (S_{r,s})_{r,s=\overline{1,q}}$ *with elements*

$$\begin{gathered} S_{r,s} = \frac{1}{f_m(Q^{F_m}(\alpha_r)) f_m(Q^{F_m}(\alpha_s))} \times \\ \times \left( \sum_{k=1}^{M} \langle p^k (a^m)^2 \rangle F_k(\min\{Q^{F_m}(\alpha_r), Q^{F_m}(\alpha_s)\}) - \right. \\ \left. - \sum_{k,l=1}^{M} \langle p^k p^l (a^m)^2 \rangle F_k(Q^{F_m}(\alpha_r)) F_l(Q^{F_m}(\alpha_s)) \right). \end{gathered} \tag{3}$$

**Quantile estimator (for Gaussian distribution)**

Let $F_1(x; t)$ be the CDF of a Gaussian distribution with the true value of a parameter $\vartheta = (\mu, \sigma)^T \in \mathbb{R}^2$.

We denote the interquartile range of standard Gaussian distribution as $\gamma := Q^{\mathcal{N}(0,1)}(3/4) - Q^{\mathcal{N}(0,1)}(1/4)$ (approximately 1.34898), and $E_\theta := \begin{pmatrix} 0 & 1 & 0 \\ -1/\gamma & 0 & 1/\gamma \end{pmatrix}$. Quantile estimator for Gaussian component is defined in [2] as

$$\hat{\vartheta}^{quant}_N := (\hat{\mu}^{quant}_N, \hat{\sigma}^{quant}_N)^T := E_\vartheta \cdot (\hat{Q}^1_N(1/4), \hat{Q}^1_N(1/2), \hat{Q}^1_N(3/4))^T. \tag{4}$$

**Corollary 1.** *(from theorem 1)*

*Let* $(\alpha_i)_{i=\overline{1,q}} = (1/4, 1/2, 3/4)^T$. *Assume that the assumptions 1-6 from theorem 1 hold.*

*Then the vector* $\sqrt{N}(\hat{\vartheta}^{quant}_N - \vartheta)$ *weakly converges to Gaussian distribution with zero mean and covariance matrix* $E_\vartheta \cdot S \cdot E^T_\vartheta$ *with* $S$ *defined by (3).*

# 4 Adaptive estimator (from GEE method)

Adaptive estimators are constructed in [1] as GEE estimators with the estimating function adapted by data to derive optimal dispersion matrices. For practical needs, it is recommended in [1] to consider a vector of some predefined parametrized functions $u(x;t) \in \mathbb{R}^R$, and choose the estimating function as a linear combination of $u(x;t)$: $B(t) \cdot u(x;t)$, where $B(t)$ is some $d$-by-$R$ matrix. Approximate adaptive estimators are obtained from pilot estimators as one-step Newton type approximate solutions of adapted estimating equations. Any $\sqrt{N}$-consistent estimator such as a moment or a quantile one can be used as the pilot estimator $\tilde{\vartheta}_N$. Thus, adaptive estimator takes form

$$\hat{\vartheta}_N^{adapt} := \tilde{\vartheta}_N - \hat{B}_N(\tilde{\vartheta}_N) \cdot \hat{u}_{1;N}(\tilde{\vartheta}_N) \tag{5}$$

where $\hat{B}_N(\tilde{\vartheta}_N)$ and $\hat{u}_{1;N}(\tilde{\vartheta}_N)$ are estimations for $B(\vartheta)$ and $\int u(x;\vartheta)F_1(dx;\vartheta)$ respectively.

Consistency and asymptotic normality of the adaptive estimator defined by (5) are demonstrated in [1].

# 5 Numerical examples

We assessed performance of the following estimators by simulations.
1. Simple estimate $\hat{\vartheta}_N^{simple}$ defined by (1) with $h(x) := (x, x^2)^T$.
2. Improved estimate $\hat{\vartheta}_N^{impr}$ defined by (2) with $h(x) := (x, x^2)^T$.
3. Quantile estimate $\hat{\vartheta}_N^{quant}$ defined by (4).
4. Adaptive estimate $\hat{\vartheta}_N^{adapt}$ defined by (5) with $\hat{\vartheta}_N^{impr}$ as a pilot.
5. Adaptive estimate $\hat{\vartheta}_N^{adapt}$ defined by (5) with $\hat{\vartheta}_N^{quant}$ as a pilot.

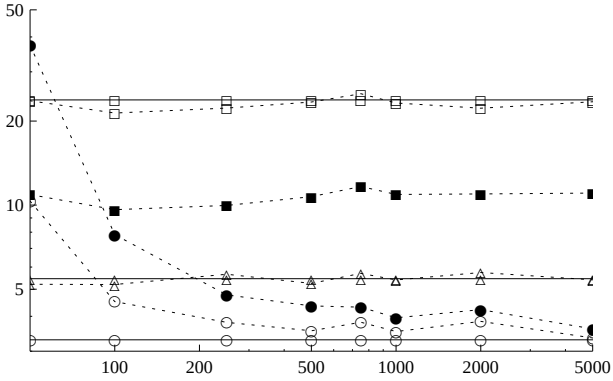
Experiments were conducted on two types of two-component mixture from Gaussian distributions with the following parameters:
Experiment 1. Component 1: $\mu = -3$, $\sigma = 1$; component 2: $\mu = 3$, $\sigma = 2$.
Experiment 2. Component 1: $\mu = 0$, $\sigma = 2$; component 2: $\mu = 1$, $\sigma = 2$.

The estimates were calculated for different sizes of a sample (value $N$): 50, 100, 250, 500, 750, 1000, 2000, 5000. The dispersion of constructed estimates was calculated from 1000 samples (for each value of $N$). The set of concentration was uniform: $p_{j;N}^1 = \{\frac{j}{N}\}_{j=1}^N$, $p_{j;N}^2 = 1 - p_{j;N}^1$.

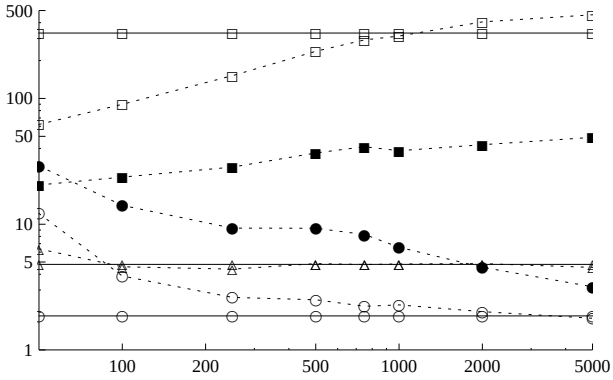
For adaptive estimate as a vector $u(x;t)$ is taken a vector from 8 functions. First 5 of them are cubic B-splines with support $(t_1 - 4t_2, t_1 + 4t_2)$ and uniform subdivision of this support into 8 intervals. The last 3 functions: 1, $(x - t_1)/t_2$, $(x - t_1)^2/t_2^2$.

The results of simulation are presented on figure 1.

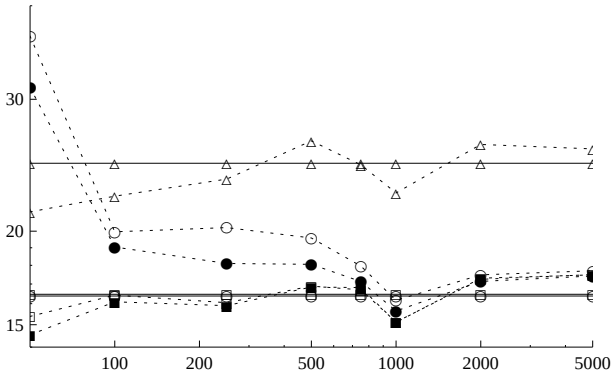
So, in our experiments the adaptive estimators outperformed the other ones in almost all cases for sample sizes larger then 100.



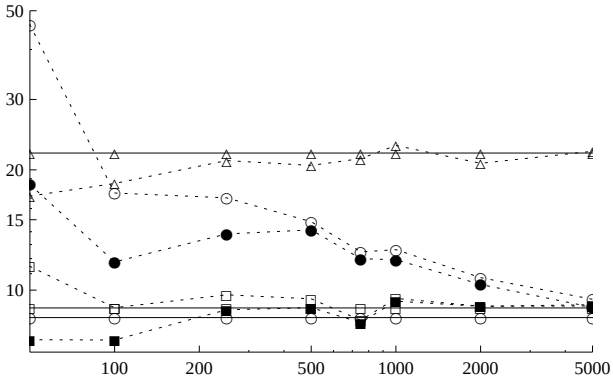

(Experiment 1, $\mu$)




(Experiment 1, $\sigma$)




(Experiment 2, $\mu$)




(Experiment 2, $\sigma$)

Figure 1: Dispersion of estimates: □ – simple estimates, ■ – improved estimates, △ – quantile estimates, • and ∘ – adaptive estimates with improved and quantile as pilot ones respectively. Asymptotic values are presented by dotted lines.

# References

[1] Maiboroda R.E., Sugakova O.V., Doronin A.V. Generalized estimating equations for mixtures with varying concentrations. *The Canadian Journal of Statistics* to appear. Published on-line http://onlinelibrary.wiley.com/doi/10.1002/cjs.11170/abstract

[2] Doronin A.V. (2012). Robust Estimates for Mixtures with Gaussian Component. *Bulletin of Taras Shevchenko National University of Kyiv. Series: Physics & Mathematics* (in Ukrainian). No. **1**, pp. 18-23.

[3] Maiboroda R.E., Kubaichuk O.O. (2005). Improved estimators for moments constructed from observations of a mixture. *Theory of Probability and Mathematical Statistics*. Vol. **70**, pp. 83-92.

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