Boundary effect for mixtures of regressions

D. D. Horbunov[1], **R. E. Maiboroda**[1]
[1]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
danielhorbunov@knu.ua, rostmaiboroda@knu.ua

Consider the model of mixtures with varying concentrations [1], where each object O_j of sample $O_1, \dots, O_n$ belongs to one of M populations (components of mixture). The true number of a component number κ_j for which O_j belongs to is unknown, but its distribution is known:

$$\mathbf{P}\left(\kappa_j = k\right) = p_{j:n}^{(k)},$$

where the probabilities $\{p_{j:n}^{(k)}\}$ are called concentrations.

For each object O_j a bivariate vector of features (X_j, Y_j) is observed. The interactions between X_j and Y_j are described in the nonparametric form

$$Y_j = g^{(\kappa_j)}(X_j) + \varepsilon_j, \tag{1}$$

where $g^{(k)}$ is an unknown regression function for k-th component of mixture, ε_j is a random error term, the distribution of ε_j can be different for different components of mixture. It is assumed that the distribution of $X_j \mid \{\kappa_j = k\}$ is absolutely continuous with probability density function $f^{(k)}$ respectively.

In order to estimate $g^{(k)}$, consider the modified Nadaraya-Watson estimator from [2]:

$$\hat{g}_{NW,n}^{(m)}(x_0) = \frac{\hat{S}_{0,1:n}^{(m)}}{\hat{S}_{0,0:n}^{(m)}} \tag{2}$$

and the modified local-linear regression estimator from [3]:

$$\hat{g}_{LLR,n}^{(m)}(x_0) = \frac{\hat{S}_{2,0:n}^{(m)}\hat{S}_{0,1:n}^{(m)} - \hat{S}_{1,1:n}^{(m)}\hat{S}_{1,0:n}^{(m)}}{\hat{S}_{2,0:n}^{(m)}\hat{S}_{0,0:n}^{(m)} - (\hat{S}_{1,0:n}^{(m)})^2}, \tag{3}$$

where the weighted sums $\hat{S}_{p,q:n}^{(m)} = \hat{S}_{p,q:n}^{(m)}(x_0)$ are defined as follows:

$$\hat{S}_{p,q:n}^{(m)}(x_0) = \frac{1}{nh}\sum_{j=1}^{n} a_{j:n}^{(m)} K\left(\frac{x_0 - X_j}{h}\right)\left(\frac{x_0 - X_j}{h}\right)^p Y_j^q,$$

$K : \mathbb{R} \to \mathbb{R}_+$ is a kernel function and $h > 0$ is a bandwidth parameter, $\{a_{j:n}^m\}$ are the minimax coefficients, described in [1].

Consider the estimation of $g^{(m)}$ at such point x_0, where $f^{(m)}$ has a jump discontinuity. In case of the homogeneous distribution of data, it is known that the local-linear regression estimator has the better asymptotic properties on the boundary of density function of the regression in comparison to the Nadaraya-Watson estimator, considering that the rate of convergence remains the same regardless of the type of point at which the regression function is estimated, see [4] and [5].

For model (1) it is empirically known that the local-linear regression estimator (3) outperforms the Nadaraya-Watson estimator (2) in terms of bias and rate of convergence to $g^{(m)}(x_0)$,

see [6]. In this talk, we present theoretical results on weak convergence of estimators (2) and (3) for the case of jump point of $f^{(m)}$.

The results are obtained with the help of the asymptotics of the vector of weighted sums. Denote $\mathbf{S}_n^{(m)} = (S_{0,0:n}^{(m)}, S_{0,1:n}^{(m)}, S_{1,0:n}^{(m)}, S_{1,1:n}^{(m)}, S_{2,0}^{(m)})^T$ the vector of weighted sums and its normalization as $\Delta_n^{(m)} = \sqrt{nh} \cdot (\mathbf{S}_n^{(m)} - \mathbf{E}[\mathbf{S}_n^{(m)}])$. Averaging operator is denoted as $\langle \mathbf{v} \rangle_n = \frac{1}{n} \sum_{j=1}^n v_j$, $\mathbf{v} = (v_1, \dots, v_n)^T \in \mathbf{R}^n$. Arithmetic operations with vectors in averaging are performed entry-wise.

Theorem 1. *Consider the following conditions:*

1. *For all $k = \overline{1, M}$ the limits exist and finite:*
$$f^{(k)}(x_0\pm) = \lim_{x \to x_0 \pm 0} f^{(k)}(x),\ g^{(k)}(x_0\pm) = \lim_{x \to x_0 \pm 0} g^{(k)}(x),$$
2. *$g^{(m)}$ is twice continuosly differentiable in the neighbourhood of x_0,*
3. *There exist $\lim\limits_{n\to\infty} \Gamma_n = \Gamma$, where $\Gamma_n = (\langle p^{(k_1)} p^{(k_2)} \rangle_n)_{k_1,k_2=1}^M$,*
4. *For all $k, k_1, k_2 = \overline{1, M}$ the limits exist and finite:*
$$\left\langle (\mathbf{a}^{(m)})^2 \mathbf{p}^{(k)} \right\rangle = \lim_{n\to+\infty} \left\langle (\mathbf{a}^{(m)})^2 \mathbf{p}^{(k)} \right\rangle_n,\ \left\langle \mathbf{a}^{(m)} p^{(k_1)} p^{(k_2)} \right\rangle = \lim_{n\to\infty} \left\langle \mathbf{a}^{(m)} p^{(k_1)} p^{(k_2)} \right\rangle_n$$
5. *$h = h_n$: $h \to 0$, $nh \to \infty$, $n \to \infty$,*
6. *$suppK(z) \subset [-A, A]$ for some $A > 0$,*
7. *$\int_{-\infty}^{\infty} z^2 K(z) dz < \infty$, $\int_{-\infty}^{\infty} z^4 (K(z))^2 dz < \infty$,*
8. *$\mathbf{E}[\varepsilon_j^4 \mid \kappa_j = k] < \infty$ are finite for all $j = \overline{1, n}$, $k = \overline{1, M}$.*

Then the following weak convergence holds:
$$\Delta_n^{(m)} \to^W N(\mathbf{0}, \Sigma^{(m)}).$$

From Theorem 1, the results on estimators (2) and (3) are obtained. Specifically, the distributions of the normalized estimators
$$n^{1/3}(\hat{g}_{NW,n}^{(m)}(x_0) - g^{(m)}(x_0)),\ n^{2/5}(\hat{g}_{LLR,n}^{(m)}(x_0) - g^{(m)}(x_0))$$
are asymptotically normal if one sets $h = Hn^{-1/3}$ and $h = Hn^{-1/5}$ respectively.

These theoretical results are demonstrated through the simulations.

1. Maiboroda R., Sugakova O. Estimation and Classification by Observations from Mixture. — Kyiv: Kyiv University Publishers, 2008, 213 p. (in Ukrainian)
2. Dychko H., Maiboroda R. A generalized Nadaraya-Watson estimator for observations obtained from a mixture. Theory Probab. Math. Stat., 2020, 100, 61-76.
3. Horbunov D., Maiboroda R. Cross-validation for local-linear regression by observations from mixture. Bull. Taras Shevchenko Natl. Univ. Kyiv., Ser. Phys. Math., 2023, 1, 37-43. (in Ukrainian)
4. Fan J. Local Linear Regression Smoothers and their minimax efficiencies, Ann. Stat., 21, 1, 1993, 196-216.
5. Ruppert D., Wand M. Multivariate Locally Weighted Least Squares Regression. Ann. Stat., 22, 3, 1994, 1346-1370.
6. Horbunov D., Maiboroda R. Consistency of local linear regression estimator for mixtures with varying concentrations, Modern Stoch. Theory Appl., 11, 3, 2024, 359-372.