Local linear regression for contaminated regression mixture

D. D. Horbunov[1], **R.E. Maiboroda**[1]

[1]Taras Shevchenko National University of Kyiv, Kyiv, Ukraine

danielhorbunov@knu.ua, rostmaiboroda@knu.ua

This talk focuses on the problem of nonparametric estimation in a mixture regression model with varying concentrations [5]. The asymptotic properties of the Nadaraya-Watson and local linear regression estimators for mixtures consisting purely of regression components were investigated in [1, 3, 4]. It remains to be established how these estimators behave in the presence of a single regression component. Below we define a nonparametric estimation problem for a such setting.

Let us consider a sample with n subjects $O_1, \dots, O_n$. Each subject O_j belongs to one of the M populations (components of mixture). The index of component $\kappa_j = \kappa(O_j)$, which the j-th object belongs to, is unknown. But the probability $p^{(m)}_{j:n} = \mathbf{P}(\kappa_j = m)$ that O_j belongs to m-th component is known for all $j = \overline{1,n}$ and $m = \overline{1,M}$. The probabilities $\{p^{(m)}_{j:n}\}$ are called the mixing probabilities (concentrations) of the components in the mixture.

For each subject O_j one observes a bivariate vector (X_j, Y_j), where $X_j = X(O_j)$ and $Y_j = Y(O_j)$ are the regressor and response respectively. It is assumed that the vectors $\{(X_j, Y_j)\}_{j=1}^n$ are mutually independent for any fixed $n \geq 1$.

Let $(X_{(m)}, Y_{(m)})$ be a random vector, which has the distribution of a (X_j, Y_j) given $\{\kappa_j = m\}$. For a fixed $k = \overline{1,M}$:

1. If $m \neq k$, there exist a joint PDF $f^{(m)}(x, y)$ for $(X_{(m)}, Y_{(m)})$,
2. Otherwise, for $m = k$ the following nonparametric regression model holds

$$Y_{(k)} = g(X_{(k)}) + \varepsilon, \tag{1}$$

where $g(x)$ is an unknown regression function and ε is a random error term. It is also assumed that there exists a PDF $f(x)$ for $X_{(k)}$. Here $X_{(k)}$ and ε are independent random variables.

The proposed model introduces non-regression components (contamination) into the mixture. This contrasts with the previously considered models of regression mixture in [1–4], which were assumed to consist entirely of the regression components. In those prior models, for all $m = \overline{1,M}$, the following regression model holds:

$$Y_{(m)} = g^{(m)}(X_{(m)}) + \varepsilon_{(m)},$$

where $g^{(m)}$ is an unknown regression function, $\varepsilon_{(m)}$ is a random error term for m-th regression component.

Based on the model defined in (1), the main task is to estimate the regression function $g(x_0)$ at a given point $x_0 \in \mathbb{R}$. For this purpose we consider the modified local-linear regression estimator (mLLRE) from [2]:

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

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

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

$K : \mathbb{R} \to [0, +\infty)$ is a kernel function and $h > 0$ is a bandwidth parameter, $\{a^{(k)}_{j:n}\}$ are the minimax coefficients, described in [5].

Let us consider a vector of weighted sums:

$$\mathbf{S}^{(k)}_n = (\hat{S}^{(k)}_{0,0:n}, \hat{S}^{(k)}_{0,1:n}, \hat{S}^{(k)}_{1,0:n}, \hat{S}^{(k)}_{1,1:n}, \hat{S}^{(k)}_{2,0:n})^T$$

and we denote the normalized version of $\mathbf{S}^{(k)}_n$ as

$$\Delta^{(k)}_n = \sqrt{nh}(\mathbf{S}^{(k)}_n - \mathbf{e}^{(k)}_n), \text{ where } \mathbf{e}^{(k)}_n = \mathbf{E}[\mathbf{S}^{(k)}_n].$$

We will show that, under certain conditions, the asymptotic normality holds for $\Delta^{(k)}_n$. This result will be applied to obtain the asymptotic distribution for the mLLRE.

[1] Dychko H., Maiboroda R., A generalized Nadaraya-Watson estimator for observations obtained from a mixture, *Theory Probab. Math. Stat.* **100** (2020), 61–76.

[2] Horbunov D., Maiboroda R., Cross-validation for local-linear regression by observations from mixture, *Bulletin of Taras Shevchenko National University of Kyiv. Physics and Mathematics* **1** (2023), 37–43.

[3] Horbunov D., Maiboroda R., Consistency of local linear regression estimator for mixtures with varying concentrations, *Modern Stoch. Theory Appl.* **11** (2024), no. 3, 359–372.

[4] Horbunov D., Maiboroda R., Asymptotic normality of local linear regression estimator for mixtures with varying concentrations, *Modern Stoch. Theory Appl.* **13** (2026), no. 1, 1–17.

[5] Maiboroda R., Sugakova O., Estimation and Classification by Observations from Mixture, Kyiv University Publishers, Kyiv, 2008, 213 pp. (in Ukrainian).