Local parametric methods in nonparametric estimation. 2.
Local parametric approach

Vladimir Spokoiny

Weierstraß-Institute
for Applied Analysis and Stochastics

October 1, 2006

Regression model

The (mean) regression model link the *explained variable* Y and the *explanatory variable* in the form

$$Y = f(X) + \varepsilon.$$

- *Observations* (X_i, Y_i) for $i = 1, \ldots, n$. Typically the Y_i's are independent. n is usually called the *sample size*.
- *Design* $X_1, \ldots, X_n$, $X_i \in \mathcal{X}$ where $\mathcal{X}$ is the design space. Usually either random or deterministic.
- *Regression function* $f(x)$ for $x \in \mathcal{X}$. The parametric case: $f(x) = f_{\boldsymbol{\theta}}(x)$ is known up to a parameter $\boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^p$.
- Errors ε_i. Mutually independent and zero mean. *Homoscedastic errors:* $\operatorname{Var} \varepsilon_i = \sigma^2$. *Heteroscedastic errors:* $\operatorname{Var} \varepsilon_i$ varies with i or with the location X_i.

Parametric M-estimation

Target of estimation - regression function $f(x)$.

Parametric model: $f(x) = f_{\boldsymbol{\theta}}(x)$.

M-estimate:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{n} M(Y_i - f_{\boldsymbol{\theta}}(X_i)).$$

- if $M(u) = u^2$, then $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{LSE}$, the least squares estimate
- if $M(u) = |u|$, then $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{LAD}$, the least absolute deviation estimate
- if $M(u) = -\log p(u)$ where $p(u)$ is the density of ε_i, then $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{MLE}$, the maximum likelihood estimate.

Regression-like model

Let $\mathcal{P} = (P_\upsilon, \upsilon \in \mathcal{U})$ be a parametric (exponential) family.
Regression-like model: Y_i are independent and the distribution of Y_i belongs to $\mathcal{P}$ where the parameter depends on X_i:

$$Y_i \sim P_{f(X_i)}, \qquad i = 1, \ldots, n.$$

The *regression function* $f(\cdot)$ identifies the distribution of $Y^{(n)}$.
For the case of the natural parametrization

$$\boldsymbol{E}[Y_i | X_i] = f(X_i).$$

Parametric modeling: $f(\cdot) = f_{\boldsymbol{\theta}}(\cdot)$. The MLE

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \ell(Y_i, f_{\boldsymbol{\theta}}(X_i))$$

where $\ell(y, \upsilon) = \log p(y, \upsilon)$ is the log-density of P_υ.

Examples. Constant and linear regression

Example (Constant regression)

Let $\theta \in \mathcal{U}$ and $f_{\boldsymbol{\theta}}(x) \equiv \theta$. Then

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \ell(Y_i, \theta) = n^{-1} \sum_{i=1}^{n} Y_i.$$

Example (Linear regression)

Let $\psi_1(x), \ldots, \psi_p(x)$ be given basis functions and $f_{\boldsymbol{\theta}}(x) = \theta_1 \psi_1(x) + \ldots + \theta_p \psi_p(x)$. Then

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell(Y_i, \Psi_i^{\top} \boldsymbol{\theta})$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\top}$ and $\Psi_i = \bigl(\psi_1(X_i), \ldots, \psi_p(X_i)\bigr)^{\top}$.

Localization

The global parametric assumption $f(x) \equiv f_{\boldsymbol{\theta}}(x)$ can be too restrictive, especially if the family $f_{\boldsymbol{\theta}}(\cdot)$ is simple (as for constant or linear regression).

Way out by local parametric assumption (LPA): suppose that this assumption is valid only approximately and in a small neighborhood of each point x.

Localization around x using the collection of weights $W = \{w_i\} = \{w_i(x)\}$:

$$\widetilde{\boldsymbol{\theta}}(x) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L(W, \boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \ell(Y_i, f_{\boldsymbol{\theta}}(X_i)) w_i(x).$$

Usually $w_i(x) = K_{\text{loc}}\big((X_i - x)/h\big)$ for a bandwidth h and a kernel K_{loc}.

Local constant regression

LPA: $f(X_i) \approx \theta$ for some θ in a neighborhood of x described by the weights $w_i = w_i(x)$.

Local estimate $\widetilde{f}(x) = \widetilde{\theta}(x)$:

$$\widetilde{f}(x) = \widetilde{\theta}(x) = \operatorname*{argmax}_{\theta} L(W, \theta) = \operatorname*{argmax}_{\theta} \sum_{i=1}^{n} \ell(Y_i, \theta) w_i .$$

In the case of an exponential family with the natural parametrization

$$\widetilde{f}(x) = \widetilde{\theta}(x) = N^{-1} \sum_{i=1}^{n} Y_i w_i \qquad \text{where} \qquad N = \sum_{i=1}^{n} w_i$$

means the local sample size.

Local linear regression

LPA: $f(X_i) \approx f_{\boldsymbol{\theta}}(X_i) = \Psi_i^{\top}\boldsymbol{\theta}$ if $w_i > 0$ for some $\boldsymbol{\theta} \in \Theta$.

Local estimate $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(x)$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ell(Y_i, \Psi_i^{\top}\boldsymbol{\theta}) w_i$$

A closed form solution only for the Gaussian contrast $\ell(y, \upsilon) = (y - \upsilon)^2$. Then

$$\widetilde{\boldsymbol{\theta}} = \Bigl(\sum_{i=1}^{n} \Psi_i^{\top}\Psi_i w_i\Bigr)^{-1} \sum_{i=1}^{n} Y_i \Psi_i w_i .$$

The value $f(x)$ is estimated as

$$\widetilde{f}(x) = f_{\widetilde{\boldsymbol{\theta}}}(x) = \Psi(x)^{\top}\widetilde{\boldsymbol{\theta}}.$$

Accuracy of local estimation in the parametric case

LPA: $f(X_i) \approx f_{\boldsymbol{\theta}}(X_i)$ if $w_i > 0$ for some $\boldsymbol{\theta} \in \Theta$.

Leads to the local estimate $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(x)$

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \ell(Y_i, f_{\boldsymbol{\theta}}(X_i)) w_i .$$

Theorem

Let the LPA be exactly fulfilled, i.e., $f(X_i) \equiv f_{\boldsymbol{\theta}^*}(X_i)$ *for* $w_i > 0$.
Then $L(W, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = \max_{\boldsymbol{\theta} \in \Theta} L(W, \boldsymbol{\theta}) - L(W, \boldsymbol{\theta}^*)$ *satisfies*

$$\boldsymbol{E}_{f(\cdot)} \bigl| L(W, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr|^r = \boldsymbol{E}_{\boldsymbol{\theta}^*} \bigl| L(W, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr|^r \le \mathfrak{R}_r .$$

Local confidence intervals:

$$\mathcal{E}(\mathfrak{z}) = \{ \boldsymbol{\theta} \in \Theta : L(W, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le \mathfrak{z} \}.$$

“Small modeling bias” condition

LPA: $f(X_i) \approx f_{\boldsymbol{\theta}}(X_i)$ if $w_i > 0$ for some $\boldsymbol{\theta} \in \Theta$.

Problems: how to measure the quality of the LPA?

A natural measure via the local *Kullback-Leibler* divergence. Define

$$\Delta(W, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{K}\bigl(f(X_i), f_{\boldsymbol{\theta}}(X_i)\bigr) \mathbf{1}(w_i > 0).$$

Theorem

Let $\boldsymbol{\theta}$ *and* $\Delta \geq 0$ *be such that* $\Delta(W, \boldsymbol{\theta}) \leq \Delta$. *Then*

$$\boldsymbol{E}_{f(\cdot)} \log\left(1 + \frac{\bigl|L(W, \widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})\bigr|^r}{\mathfrak{R}_r}\right) \leq \Delta + 1.$$

Interpretation: the local parametric approach applies as long as the SMB holds.

Problem of local adaptive estimation

Let $W^{(k)} = \{w_i^{(k)}\}$, $k = 1, \ldots, K$, be an ordered collection of localizing schemes for a fixed x.

Usually $w_i^{(k)} = K_{\mathrm{loc}}\bigl((X_i - x)/h_k\bigr)$ for a giving ordered set of bandwidths $h_1 < h_2 < \ldots < h_K$.

Leads to a growing local sample size $N_k = \sum w_i^{(k)}$ and decreasing variability of the $\widetilde{\boldsymbol{\theta}}_k$.

$$
\begin{array}{ccccccc}
W^{(1)} & \subset & W^{(2)} & \subset & \ldots & \subset & W^{(K)} \\
\downarrow & & \downarrow & & & & \downarrow \\
\widetilde{\boldsymbol{\theta}}_1 & & \widetilde{\boldsymbol{\theta}}_2 & & \ldots & & \widetilde{\boldsymbol{\theta}}_K \\
\downarrow & & \downarrow & & & & \downarrow \\
N_1 & < & N_2 & < & \ldots & < & N_K
\end{array}
$$

Aim: to build an estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}(x)$ which behaves as good as the best in the family $\widetilde{\boldsymbol{\theta}}_k$.

Local model selection (LMS) procedure. Idea

For a given x and a set $W^{(1)} \subset W^{(2)} \subset \ldots \subset W^{(K)}$.

Local Model Selection Problem: select the largest scheme $W^{(k)}$ with the largest N_k for which the SMB still holds.

Idea: sequential test of the hypothesis of local homogeneity $f(X_i) = f_{\boldsymbol{\theta}}(X_i)$ for $w_i^{(k)} > 0$.

If the hypothesis holds for $W^{(k)}$, the value $\boldsymbol{\theta}$ belongs with the high probability to the confidence set

$$\mathcal{E}_k = \mathcal{E}_k(\mathfrak{z}) = \{\boldsymbol{\theta} \in \Theta : L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k, \boldsymbol{\theta}) \le \mathfrak{z}\}.$$

$\widetilde{\boldsymbol{\theta}}_k$ is accepted if it belongs to all confidence sets $\mathcal{E}_l$ for $l < k$.

LMS procedure. Formal description

- Start with $\widehat{\boldsymbol{\theta}}_1 = \widetilde{\boldsymbol{\theta}}_1$.
- for $k \geq 2$, $\widetilde{\boldsymbol{\theta}}_k$ is accepted and $\widehat{\boldsymbol{\theta}}_k = \widetilde{\boldsymbol{\theta}}_k$ if $\widetilde{\boldsymbol{\theta}}_{k-1}$ was accepted and

  $$L\bigl(W^{(l)}, \widetilde{\boldsymbol{\theta}}_l, \widetilde{\boldsymbol{\theta}}_k\bigr) \leq \mathfrak{z}_l \,, \qquad l = 1, \ldots, k-1.$$

  Otherwise $\widehat{\boldsymbol{\theta}}_k = \widehat{\boldsymbol{\theta}}_{k-1}$.

$\widehat{\boldsymbol{\theta}}_k$ is the latest accepted estimate after first k steps.

The adaptive estimate $\widehat{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}}_K$ is the latest accepted estimate among $\widetilde{\boldsymbol{\theta}}_k$.

LMS procedure. Parameters

To run the procedure, one has to specify:

- Set of localizing schemes (the bandwidths h_k and the kernel K_{loc})
- the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$.

The localizing schemes $W^{(k)}$ are assumed to be given. The only condition to be verified that the local sample size $N_k = \sum_i w_i^{(k)}$ grows geometrically with k .

The critical values $\mathfrak{z}_k$ are selected to provide the prescribed performance of the method in the parametric situation:

$$\sup_{\boldsymbol{\theta}^* \in \Theta} \boldsymbol{E}_{\boldsymbol{\theta}^*} \bigl| L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\theta}}_k) \bigr|^r \leq \alpha \mathfrak{R}_r \, .$$

Sequential choice of critical values

The parameters $\mathfrak{z}_k$ have to fulfill

$$\sup_{\boldsymbol{\theta}^* \in \Theta} \boldsymbol{E}_{\boldsymbol{\theta}^*} \bigl| L(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\theta}}_k) \bigr|^r \le \alpha \mathfrak{R}_r \,, \qquad k = 2, \ldots, K. \tag{1}$$

In total $K-1$ conditions to fix $K-1$ parameters. The sensitivity to deviations from local homogeneity is important. Therefore, we aim to select the minimal $\mathfrak{z}_k$'s providing (1).

Sequential procedure.

Start with $\mathfrak{z}_1$ letting $\mathfrak{z}_2 = \ldots = \mathfrak{z}_{K-1} = \infty$. Leads to the estimates $\widehat{\theta}_t^{(k)}(\mathfrak{z}_1)$ for $k = 2, \ldots, K$. The value $\mathfrak{z}_1$ is selected as the minimal one for which

$$\boldsymbol{E}_{\theta^*} \bigl| L\bigl(W^{(k)}, \widetilde{\boldsymbol{\theta}}_k, \widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1)\bigr) \bigr|^r \le \frac{\alpha \mathfrak{r}_r}{K-1} \,, \qquad k = 2, \ldots, K. \tag{2}$$

Such a value exists because the choice $\mathfrak{z}_1 = \infty$ leads to $\widehat{\boldsymbol{\theta}}_k(\mathfrak{z}_1) = \widetilde{\boldsymbol{\theta}}_k$ for all k.

Sequential choice of critical values. 2

Suppose $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ have been already fixed.

We set $\mathfrak{z}_k = \ldots = \mathfrak{z}_{K-1} = \infty$ and fix $\mathfrak{z}_k$ leading to the set of parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_k, \infty, \ldots, \infty$ and the estimates $\widehat{\boldsymbol{\theta}}_m(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)$ for $m = k+1, \ldots, K$

We select $\mathfrak{z}_k$ as the minimal value which fulfills

$$\boldsymbol{E}_{\theta^*} \bigl| L\bigl(\widetilde{\boldsymbol{\theta}}_l, \widehat{\boldsymbol{\theta}}_l(\mathfrak{z}_1, \ldots, \mathfrak{z}_k)\bigr) \bigr|^r \le \frac{k \alpha \mathfrak{r}_r}{K-1}, \qquad l = k+1, \ldots, K. \tag{3}$$