Foundations and Applications of Modern Nonparametric Statistics

Vladimir Spokoiny and Wolfgang Härdle

Weierstraß-Institute
for Applied Analysis and Stochastics

and Institute of Statistics and Economics,
Humboldt-Universität zu Berlin

October 4, 2006

Basics

Statistics is understanding data by modeling it.

Data $Y^{(n)} = (Y_1, \ldots, Y_n)$ usually *random*.

$\boldsymbol{P} = \mathcal{L}(Y^{(n)})$, the *unknown* joint distribution.

Statistical problem: to infer on $\boldsymbol{P}$ from the data $Y^{(n)}$.

Parametric modeling:

$$\boldsymbol{P} = \boldsymbol{P}_{\boldsymbol{\theta}} \in (\boldsymbol{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subset \boldsymbol{R}^p).$$

Nonparametric modeling: the parametric assumption is not fulfilled, or, equivalently, $p = \infty$.

Example

$Y_1 = 1.2, Y_2 = 2.3, Y_3 = 0.6, \ldots, Y_{20} = -1.5$.

$n = 20$.

$Y^{(n)} = (Y_1, \ldots, Y_n)$.

Parametric modeling (for single observation): $P = P_\theta$, P_θ with pdf $\varphi(y - \theta)$.

Parametric modeling (for n observations): $\boldsymbol{P} = \boldsymbol{P}_\theta$, $\boldsymbol{P}_\theta$ with pdf $\prod_{i=1}^{n} \varphi(y_i - \theta)$.

Example

$$\varphi(y) = (2\pi)^{-1/2} \exp\bigl(-y^2/2\bigr)\,.$$

Shift model: P_θ is given by $\varphi(y-\theta)\,,\ p=1\,.$

Shift scale model: $P_{\boldsymbol{\theta}}$ is given by $\sigma^{-1}\varphi\bigl\{\sigma^{-1}(y-\mu)\bigr\}\,,$
$\boldsymbol{\theta} = (\mu,\sigma)\,,\ p=2\,.$

Nonparametric modeling

$Y^{(n)} = (Y_1, \ldots, Y_n)$ with pdf f from some function space, e.g.

$$f \in L_2 = \{f : \|f\|_2^2 < \infty, \int f = 1\}.$$

Fourier expansion:

$$f(y) = \sum_{j=1}^{\infty} \theta_j \, \psi_j(y),$$

$\{\psi_j\}$ ONB of the function space, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots)$, $p = \infty$.

Parametric estimation. Maximum Likelihood approach

A parametric model: $Y^{(n)} \sim \boldsymbol{P}_{\boldsymbol{\theta}^*}$ for some $\boldsymbol{\theta}^* \in \Theta$.

Let $\boldsymbol{P}_{\boldsymbol{\theta}} \ll \boldsymbol{P}$ for some measure $\boldsymbol{P}$ for all $\boldsymbol{\theta} \in \Theta$. Define the log-likelihood

$$L(\boldsymbol{\theta}) = \log \frac{d\boldsymbol{P}_{\boldsymbol{\theta}}}{d\boldsymbol{P}}(Y^{(n)}),$$

For some $\boldsymbol{\theta}^\circ \in \Theta$, *the (log)-likelihood ratio* is

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) = \log \frac{d\boldsymbol{P}_{\boldsymbol{\theta}}}{d\boldsymbol{P}_{\boldsymbol{\theta}^\circ}}(Y^{(n)})$$

Maximum likelihood estimate (MLE) $\widetilde{\boldsymbol{\theta}}$ is the point of maximum of $L(\boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ).$$

Minimum Contrast Estimate

Let $\mathcal{C}(\boldsymbol{\theta}) = \mathcal{C}(Y^{(n)}, \boldsymbol{\theta})$ be a *contrast function*.

$$\widetilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}}\, \mathcal{C}(\boldsymbol{\theta}).$$

Example

The choice $\mathcal{C}(\boldsymbol{\theta}) = -L(\boldsymbol{\theta})$ leads back to the MLE.

Minimum Contrast Estimate. Example

$P_\theta =$ double exponential with pdf $\frac{1}{2}\exp\bigl(-|y|\bigr)$

Likelihood function $2^{-n}\prod_{i=1}^{n}\exp\bigl(-|Y_i-\theta|\bigr)$.

Log-likelihood $L(\theta) = -\sum_{i=1}^{n}|Y_i-\theta| - n\log 2$.

$$\begin{aligned} \mathcal{C}(\theta) &= \sum_{i=1}^{n}|Y_i-\theta| \qquad \Bigl[+n\log 2\Bigr] \\ L(\theta) &= -\mathcal{C}(\theta) \end{aligned}$$

Minimum Contrast Estimate. Example

Example

Least Squares Estimate (LSE) and Least Absolute Deviation (LAD) in the parametric regression model $\boldsymbol{E}(Y_i|X_i) = f(X_i, \boldsymbol{\theta})$:

$$\widetilde{\boldsymbol{\theta}}_{LSE} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \big\{ Y_i - f(X_i, \boldsymbol{\theta}) \big\}^2$$

$$\widetilde{\boldsymbol{\theta}}_{LAD} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \big| Y_i - f(X_i, \boldsymbol{\theta}) \big|$$

Some examples of parametric families. "Gaussian shift"

Let $Y_1, \ldots, Y_n$ be i.i.d. and follow

$$Y_i = \theta^* + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with known variance σ^2.

Then

$$\begin{aligned} L(\theta) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \theta)^2. \\ \widetilde{\theta} &= \operatorname*{argmax}_{\theta} L(\theta) = n^{-1}\sum_{i=1}^{n} Y_i = \theta^* + \sigma n^{-1/2}\xi\,, \qquad (1) \end{aligned}$$

where ξ is standard normal.

"Gaussian shift". Risk and confidence interval

The MLE $\widetilde{\theta}$ fulfills $\widetilde{\theta} = \theta^* + \sigma n^{-1/2}\xi$.

For any $r > 0$

$$\boldsymbol{E}_{\theta^*}\left|\widetilde{\theta} - \theta^*\right|^r = \sigma^r n^{-r/2}\mathfrak{c}_r$$

with $\mathfrak{c}_r = \boldsymbol{E}|\xi|^r$. E.g. $\mathfrak{c}_1 = \sqrt{2/\pi}$, $\mathfrak{c}_2 = 1$ and $\mathfrak{c}_3 = 2\mathfrak{c}_1 = \sqrt{8/\pi}$, $\mathfrak{c}_4 = 3$.

If z_α is such that $P(|\xi| \leq z_\alpha) = 1 - \alpha$ with some $\alpha \in (0, 1)$, then

$$\mathcal{E}_\alpha = [\widetilde{\theta} - \sigma n^{-1/2} z_\alpha, \widetilde{\theta} + \sigma n^{-1/2} z_\alpha], \tag{2}$$

is a α-confidence interval for the parameter θ^*.

Confidence set

A confidence set (CS) to level α is a random set $\mathcal{E}_\alpha$ such that

$$\boldsymbol{P}_{\boldsymbol{\theta}^*}\left(\mathcal{E}_\alpha \ni \boldsymbol{\theta}^*\right) \geq 1-\alpha.$$

The $\mathcal{E}_\alpha$ from (2) is a CS to level α:

$$\begin{aligned}\boldsymbol{P}_{\theta^*}\left(\mathcal{E}_\alpha \ni \theta^*\right) &= \boldsymbol{P}_{\theta^*}\left(\widetilde{\theta} - \sigma n^{-1/2} z_\alpha \leq \theta^* \leq \widetilde{\theta} + \sigma n^{-1/2} z_\alpha\right) \\ &= \boldsymbol{P}_{\theta^*}\left(\sqrt{n}|\widetilde{\theta} - \theta^*|/\sigma \leq z_\alpha\right) = 1-\alpha\end{aligned}$$

since (1) yields $\xi = \sqrt{n}|\widetilde{\theta} - \theta^*|/\sigma$.

"Gaussian shift". Fitted likelihood

For any θ

$$L(\widetilde{\theta}, \theta) = L(\widetilde{\theta}) - L(\theta) = n\sigma^{-2}(\widetilde{\theta} - \theta)^2/2. \tag{3}$$

Since $\widetilde{\theta} = \theta^* + \sigma n^{-1/2}\xi$,

$$2L(\widetilde{\theta}, \theta^*) = n\sigma^{-2}(\widetilde{\theta} - \theta^*)^2 = \xi^2 \sim \chi^2_1$$

If $\mathfrak{z}_\alpha$ is the α-quantile of χ^2_1 with $P(\xi^2 > \mathfrak{z}_\alpha) = \alpha$, then

$$\mathcal{E}_\alpha = \{u : 2L(\widetilde{\theta}, u) \leq \mathfrak{z}_\alpha\} \tag{4}$$

is again an α-CS, but this time likelihood based.

Fitted likelihood "Gaussian shift". Details

The formula $L(\widetilde{\theta}, \theta) = n\sigma^{-2}(\widetilde{\theta} - \theta)^2/2$ can be seen as follows:

$$\begin{aligned} L(\widetilde{\theta}) - L(\theta) &= -(2\sigma^2)^{-1} \sum_{i=1}^{n} (Y_i - \widetilde{\theta})^2 + (2\sigma^2)^{-1} \sum_{i=1}^{n} (Y_i - \theta)^2 \\ &= n\sigma^{-2}(\widetilde{\theta} - \theta)^2/2 - \sigma^{-2} \sum_{i=1}^{n} (Y_i - \widetilde{\theta})\,(\widetilde{\theta} - \theta). \end{aligned}$$

The last sum is equal to zero by definition of $\widetilde{\theta}$.

Bernoulli model

Let $Y_1, \ldots, Y_n$ be i.i.d. Bernoulli r.v.'s satisfying $\boldsymbol{P}(Y_i = 1) = \theta$ and $\boldsymbol{P}(Y_i = 0) = 1 - \theta$. Then

$$\begin{aligned} L(\theta) &= \log \prod_{i=1}^{n} \theta^{Y_i}(1-\theta)^{1-Y_i} = \log\theta \sum_i Y_i + \log(1-\theta) \sum_i (1 - Y_i) \\ &= S \log \frac{\theta}{1-\theta} + n \log(1-\theta), \end{aligned}$$

where $S = Y_1 + \ldots + Y_n$. Hence,

$$\widetilde{\theta} = S/n, \quad \text{and} \quad L(\widetilde{\theta}, \theta) = n\widetilde{\theta} \log \frac{\widetilde{\theta}}{\theta} + n(1-\widetilde{\theta}) \log \frac{1-\widetilde{\theta}}{1-\theta} = n\mathcal{K}(\widetilde{\theta}, \theta)$$

where $\mathcal{K}(\theta, \theta') = \theta \log(\theta/\theta') + (1-\theta) \log\{(1-\theta)/(1-\theta')$ means the Kullback-Leibler divergence for the Bernoulli law.

Kullback-Leibler divergence

Kullback-Leibler (KL) divergence measures a "distance" between two distributions:

$$\mathcal{K}(P, Q) = \boldsymbol{E}_P\left\{\log\left(\frac{dP}{dQ}\right)\right\}.$$

In terms of parametric model $\boldsymbol{P}_{\boldsymbol{\theta}}$:

$$\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}') = \boldsymbol{E}_{\boldsymbol{\theta}}\left\{\log\left(\frac{d\boldsymbol{P}_{\boldsymbol{\theta}}}{d\boldsymbol{P}_{\boldsymbol{\theta}'}}\right)\right\}.$$

With pdf $p(y, \boldsymbol{\theta})$:

$$\mathcal{K}(\boldsymbol{\theta}, \boldsymbol{\theta}') = E_{\boldsymbol{\theta}}\left\{\log\frac{p(y, \boldsymbol{\theta})}{p(y, \boldsymbol{\theta}')}\right\} = E_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}, \boldsymbol{\theta}'), \qquad \ell(\boldsymbol{\theta}, \boldsymbol{\theta}') = \log\frac{p(y, \boldsymbol{\theta})}{p(y, \boldsymbol{\theta}')}.$$

Note that

$$\boldsymbol{E}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\theta}') = nE_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}, \boldsymbol{\theta}')$$

Poisson model

Let $Y_1, \ldots, Y_n$ be i.i.d. Poisson r.v.'s satisfying $\boldsymbol{P}(Y_i = m) = \theta^m e^{-\theta}/m!$ for $m = 0, 1, 2, \ldots$. Then

$$\begin{aligned} L(\theta) &= \log \prod_{i=1}^{n} \theta^{Y_i} e^{-\theta}/Y_i! = \log\theta \sum_{i=1}^{n} Y_i - \theta - \log(Y_i!) \\ &= S\log\theta - n\theta + R, \end{aligned}$$

where $S = Y_1 + \ldots + Y_n$ and $R = \sum_{i=1}^{n} \log(Y_i!)$. Therefore,

$$\widetilde{\theta} = S/n, \quad \text{and} \quad L(\widetilde{\theta}, \theta) = n\widetilde{\theta}\log(\widetilde{\theta}/\theta) - n(\widetilde{\theta} - \theta) = n\mathcal{K}(\widetilde{\theta}, \theta)$$

where $\mathcal{K}(\theta, \theta') = \theta\log(\theta/\theta') - (\theta - \theta')$ means the Kullback-Leibler divergence for the Poisson law.

Poisson model. Details

$$L(\widetilde{\theta}, \theta) = n\widetilde{\theta} \log(\widetilde{\theta}/\theta) - n(\widetilde{\theta} - \theta)$$

$$\begin{aligned}
n\mathcal{K}(\theta, \theta') &= \boldsymbol{E}_\theta L(\theta, \theta') \\
&= \boldsymbol{E}_\theta L(\theta) - \boldsymbol{E}_\theta L(\theta') \\
&= \boldsymbol{E}_\theta \Big[S \log \theta - n\theta \Big] - \boldsymbol{E}_\theta \Big[S \log \theta' - n\theta' \Big] \\
&= \boldsymbol{E}_\theta S \log(\theta/\theta') - n(\theta - \theta') \\
&= n\theta \log(\theta/\theta') - n(\theta - \theta')
\end{aligned}$$

Exponential model

Let $Y_1, \ldots, Y_n$ be i.i.d. exponential r.v.'s with parameter $\theta > 0$: $\boldsymbol{P}(Y_i > t) = e^{-t/\theta}$. Then

$$L(\theta) = -n \log \theta - \sum_{i=1}^{n} Y_i/\theta = -S/\theta - n \log \theta,$$

where $S = Y_1 + \ldots + Y_n$. Therefore

$$\widetilde{\theta} = S/n, \quad \text{and} \quad L(\widetilde{\theta}, \theta) = -n(1 - \widetilde{\theta}/\theta) - n \log(\widetilde{\theta}/\theta) = n\mathcal{K}(\widetilde{\theta}, \theta)$$

where $\mathcal{K}(\theta, \theta') = \theta/\theta' - 1 - \log(\theta/\theta')$ means the Kullback-Leibler divergence for the exponential law.

Volatility model

Let $\xi_1, \ldots, \xi_n$ be i.i.d. $\mathcal{N}(0, \theta)$ r.v.'s and we observe $Y_i = \xi_i^2$. Then

$$L(\theta) = -\frac{n}{2}\log(2\pi\theta) - \sum_{i=1}^{n} Y_i/(2\theta) = -\frac{n}{2}\log(2\pi\theta) - S/(2\theta),$$

where $S = Y_1 + \ldots + Y_n$. Therefore,

$$\widetilde{\theta} = S/n \quad \text{and} \quad L(\widetilde{\theta}, \theta) = -\frac{n}{2}\log(\widetilde{\theta}/\theta) - \frac{n}{2}(1 - \widetilde{\theta}/\theta) = n\mathcal{K}(\widetilde{\theta}, \theta)$$

where $\mathcal{K}(\theta, \theta') = 0.5(\theta/\theta' - 1) - 0.5\log(\theta/\theta')$ means the Kullback-Leibler divergence for the two zero mean normal laws with the variance θ' and θ.

Exponential family

Means that all measures P_θ have density functions $p(y,\theta) = dP_\theta/dP(y)$ of the form

$$p(y,\theta) = p(y)e^{yC(\theta)-B(\theta)}.$$

Here $C(\theta)$ and $B(\theta)$ are some given nondecreasing functions on Θ and $p(y)$ is some nonnegative function on $\mathcal{Y}$.

The *natural* parametrization means the relation $E_\theta Y = \theta$. This and the identity $\boldsymbol{E}p(y,\theta) \equiv 1$ yield

$$B'(\theta) = \theta C'(\theta).$$

Moreover, the Kullback-Leibler divergence $\mathcal{K}(\theta,\theta') = E_\theta \log\{p(Y,\theta)/p(Y,\theta')\}$ for $\theta,\theta' \in \Theta$ satisfies

$$\mathcal{K}(\theta,\theta') = \theta\{C(\theta) - C(\theta')\} - \{B(\theta) - B(\theta')\}. \qquad (5)$$

MLE for exponential family model

The density $p(y, \theta) = p(y)e^{yC(\theta)-B(\theta)}$ leads to the log-likelihood

$$L(\theta) \quad = \quad \sum_{i=1}^{n} \log p(Y_i, \theta) = SC(\theta) - nB(\theta) + R \qquad (6)$$

where

$$S = \sum_{i=1}^{n} Y_i, \qquad R = \sum_{i=1}^{n} \log p(Y_i).$$

The estimating equation $L'(\theta) = 0$ for $\widetilde{\theta}$ and the identity $B'(\theta) \equiv \theta C'(\theta)$ result in

$$\widetilde{\theta} = S/n = n^{-1} \sum_{i=1}^{n} Y_i \quad \text{and} \quad L(\widetilde{\theta}, \theta) = L(\widetilde{\theta}) - L(\theta) = n\mathcal{K}(\widetilde{\theta}, \theta)\,.$$

Exponential bound for the fitted likelihood. Exponential family

Theorem (Polzehl and Spokoiny (2005))

Let (P_θ) *be an exponential family. Then for any* $\mathfrak{z} > 0$

$$\boldsymbol{P}_{\theta^*}\left(L(\widetilde{\theta}, \theta^*) > \mathfrak{z}\right) \leq 2e^{\mathfrak{z}}.$$

Moreover, for any $r > 0$

$$\boldsymbol{E}_{\theta^*} L^r(\widetilde{\theta}, \theta^*) = n^r \boldsymbol{E}_{\theta^*} \mathcal{K}^r(\widetilde{\theta}, \theta^*) \leq \mathfrak{r}_r\,,$$

where $\mathfrak{r}_r = 2r \int_{\mathfrak{z} \geq 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} = 2r\Gamma(r)\,.$

θ^* maximizes $\boldsymbol{E}_{\theta^*} L(\theta, \theta^*)$ over θ and $\boldsymbol{E}_{\theta^*} L(\theta^*, \theta^*) = 0\,.$

$\widetilde{\theta}$ maximizes $L(\theta, \theta^*)$ over θ, thus $L(\widetilde{\theta}, \theta^*) \geq 0\,.$ By Theorem 3 $L(\widetilde{\theta}, \theta^*)$ is stochastically bounded.

A general exponential bound for the fitted likelihood

Theorem (Golubev and Spokoiny (2006))

Let $Y^{(n)}$ *be from a parametric family* $(\boldsymbol{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta)$. *Then under some regularity conditions for any* $r > 0$

$$\boldsymbol{E}_{\theta^*} L^r(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq \mathfrak{R}_r \,,$$

where $\mathfrak{R}_r$ *depends on the parametric family only.*
Moreover, if $\mathcal{V}^2(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathrm{Var}\, L(\boldsymbol{\theta}, \boldsymbol{\theta}')$, *then*

$$\boldsymbol{E}_{\boldsymbol{\theta}^*} \mathcal{V}^{2r}(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \leq \mathfrak{R}_r^* \,.$$

Some corollaries

In the regular case,

$$L(\widetilde{\theta}, \theta^*) \approx \frac{n}{2}(\widetilde{\theta} - \theta^*)^\top I(\theta^*)(\widetilde{\theta} - \theta^*)$$

where $I(\theta^*)$, the Fisher information matrix, and the theorem yields root-n consistency:

$$E_{\theta^*}^{1/r} \left| \sqrt{I(\theta^*)}(\widetilde{\theta} - \theta^*) \right|^r \leq C/\sqrt{n}.$$

Confidence sets: define

$$\mathcal{E}(\mathfrak{z}) = \{\theta : L(\widetilde{\theta}, \theta^*) \leq \mathfrak{z}\}.$$

If $\mathfrak{z}$ is sufficiently large then

$$P_{\theta^*}\left(\mathcal{E}(\mathfrak{z}) \not\ni \theta^*\right) \approx 0.$$

Advantages and problems of the parametric modeling

1. Well developed algorithms
2. Nice asymptotic theory. Root-n consistence and asymptotic normality of the estimator $\widehat{\boldsymbol{\theta}}$ under mild regularity assumptions.
3. Good in-sample properties.

Problem:

1. The parametric structure is crucial. If the parametric assumption is violated, the MLE estimator $\widetilde{\boldsymbol{\theta}}$ is often completely misspecified.

Aim: To extend the parametric approach and methods to the situation when the parametric assumption is not precisely fulfilled.