



Weierstraß-Institut für Angewandte Analysis und Stochastik

Vladimir Spokoiny

## Foundations and Applications of Modern Nonparametric Statistics



Leibniz  
Gemeinschaft

---

Mohrenstr. 39, 10117 Berlin  
[www.wias-berlin.de/spokoiny](http://www.wias-berlin.de/spokoiny)

[spokoiny@wias-berlin.de](mailto:spokoiny@wias-berlin.de)  
October 9, 2009

## Notations

---

The 3 letter code of this course is:

**MNS**, meaning **modern nonparametric statistics**.

Software code for the examples is available on demand.

## Notations

$\mathbf{Y}$	$(Y_1, \dots, Y_n)$ data sample
$P$	distribution of a single observation
$P_\theta$	parametric distribution of a single observation
$IP$	distribution of the sample $\mathbf{Y}$
$IP_\theta$	parametric distribution of the sample $\mathbf{Y}$
$E$	Expectation operator
$L(\theta)$	$= \log \frac{dP_\theta}{dP}(\mathbf{Y})$ , log-likelihood for $IP_\theta$
$L(\theta, \theta')$	$= L(\theta) - L(\theta')$ , log-likelihood ratio of $IP_\theta$ w.r.t. $IP_{\theta'}$
$\mathcal{N}$	normal distribution

## Notations

---

$f(x)$	regression function
$f(x, \theta)$	parametric regression function
EF	exponential family
$\ell(y, v)$	$= \log p(y, v)$ , log density of $P_v$
$\mathcal{K}(P, Q)$	Kullback-Leibler divergence between measures $P, Q$
$\mathcal{K}(\theta, \theta')$	Kullback-Leibler divergence between $P_\theta$ and $P_{\theta'}$
$I(\theta)$	Fisher information matrix at $\theta$

## Notations

---

$\theta$	parameter one dimensional
$\boldsymbol{\theta}$	parameter multi dimensional
$\boldsymbol{\theta}^*$	true parameter $f(\cdot) \equiv f(\cdot, \boldsymbol{\theta}^*)$
<i>LPA</i>	local parametric approximation
$W$	$\{w_i\}$ weighting scheme
$\tilde{\boldsymbol{\theta}}$	$= \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$ , local ML estimate for $W$
$\mathfrak{c}_r$	$= E \xi ^{2r}$ , risk bound for Gaussian shift model
$\mathfrak{r}_r$	risk bound for EF
$\mathfrak{R}_r$	risk bound in a parametric model

## Notations

$W^{(k)}$	$k$ -th weighting scheme
$\tilde{\theta}_k$	estimate for $W^{(k)}$
$\mathfrak{z}_k$	$k$ -th critical value
$\hat{\theta}_k$	adaptive estimate after $k$ steps
$\hat{\theta}$	final adaptive estimate
$\hat{k}$	selected model
$k^\circ$	“oracle choice”
$\Delta(W, \theta)$	modeling bias
SMB	“small modeling bias” condition

## Overview

---

▶ Log-likelihood:

$$L(\boldsymbol{\theta}) = \log(dP_{\boldsymbol{\theta}}/d\mathbb{P})$$

▶ Maximum likelihood estimate:

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}).$$

▶ Fitted (log-)likelihood:

$$L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \max_{\boldsymbol{\theta}'} L(\boldsymbol{\theta}') - L(\boldsymbol{\theta}).$$

## Parametric risk bound

- Parametric risk bound:

$$\mathbb{E}_{\theta^*} |L(\tilde{\theta}, \theta^*)|^r \leq \mathfrak{R}_r(\theta^*) \leq \mathfrak{R}_r.$$

- Gaussian shift (GS) case:  $Y_i \sim \mathcal{N}(0, \sigma^2)$  i.i.d.:

$$L(\tilde{\theta}, \theta^*) = \frac{n(\tilde{\theta} - \theta^*)^2}{2\sigma^2}, \quad \mathbb{E}_{\theta^*} \left| \frac{n(\tilde{\theta} - \theta^*)^2}{2\sigma^2} \right|^r = \mathfrak{c}_r \equiv E|\xi|^{2r}$$

- Exponential family (EF) case:  $Y_i \sim P_\theta \in \mathcal{P}$ :

$$L(\tilde{\theta}, \theta^*) = n\mathcal{K}(\tilde{\theta}, \theta^*), \quad \mathbb{E}_{\theta^*} |n\mathcal{K}(\tilde{\theta}, \theta^*)|^r \leq \mathfrak{r}_r \equiv 2r\Gamma(r).$$



## Local parametric approach

- ▶ **Regression-like models:**  $Y_i \sim P_{f(X_i)} \in \mathcal{P} = (P_v, v \in \mathcal{U})$ .
- ▶ **Parametric modeling:**  $f(\cdot) = f(\cdot, \boldsymbol{\theta})$ .

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_i \ell\{Y_i, f(X_i, \boldsymbol{\theta})\}.$$

- ▶ **Local parametric assumption (LPA):**  $W = (w_i)$ , a localizing scheme,  $f(X_i) \approx f(X_i, \boldsymbol{\theta})$  for  $w_i > 0$ .
- ▶ **Local parametric estimation:**

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} L(W, \boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta}} \sum_i \ell\{Y_i, f(X_i, \boldsymbol{\theta})\} w_i.$$

## “Small modeling bias” condition

- ▶ “Small modeling bias” condition: for some  $\theta$

$$\Delta(W, \theta) = \sum_i \mathcal{K}\{f(X_i), f(X_i, \theta)\} \mathbf{1}(w_i > 0) \leq \Delta.$$

- ▶ “Information theoretic bound”: for any  $\zeta \sim \mathcal{F}_W = \sigma\{Y_i \mathbf{1}(w_i > 0), i = 1, \dots, n\}$

$$\mathbb{E} \log(1 + \zeta) \leq \Delta + \mathbb{E}_{\theta} \zeta.$$

- ▶ Risk bound under SMB:

$$\mathbb{E} \log\left(1 + \frac{|L(\tilde{\theta}, \theta)|^r}{\mathfrak{R}_r(\theta)}\right) \leq \Delta + 1.$$

## Local model selection

Set-up: given an **ordered set** of localizing schemes

$$\begin{array}{ccccccc}
 W^{(1)} & \subset & W^{(2)} & \subset & \dots & \subset & W^{(K)} \\
 \downarrow & & \downarrow & & & & \downarrow \\
 \tilde{\theta}_1, N_1 & & \tilde{\theta}_2, N_2 & & \dots & & \tilde{\theta}_K, N_K
 \end{array}$$

with  $W^{(k)} = \{w_i^{(k)}\}$ ,  $N_k = \sum_i w_i^{(k)}$ ,  $\tilde{\theta}_k = \operatorname{argmax}_{\theta} L(W^{(k)}, \theta)$ .

► **Local model selection:**

$$\hat{k} = \max\{k : L(W^{(\ell)}, \tilde{\theta}_\ell, \tilde{\theta}_m) \leq \beta \ell \quad \forall \ell < m \leq k\}, \quad \hat{\theta} = \tilde{\theta}_{\hat{k}}.$$

► **Restricted procedure:**

$$\hat{\theta}_k = \tilde{\theta}_{\min\{\hat{k}, k\}} \quad k \leq K.$$

## “Propagation” condition

- Parameters (“critical values”)  $\mathfrak{z}_1, \dots, \mathfrak{z}_K$ . Selected by the “propagation” condition:

$$\mathbb{E}_{\theta^*} |L(W^{(k)}, \tilde{\theta}_k, \hat{\theta}_k)|^r \leq \rho \mathfrak{R}_r(\theta^*).$$

- “Propagation” condition for **local constant GR**:

$$\mathbb{E}_0 |0.5 N_k (\tilde{\theta}_k - \hat{\theta}_k)^2|^r \leq \rho \mathfrak{c}_r.$$

- “Propagation” condition for **local constant EF**:

$$\mathbb{E}_{\theta^*} |N_k \mathcal{K}(\tilde{\theta}_k, \hat{\theta}_k)|^r \leq \rho \mathfrak{r}_r(\theta^*).$$

## “Propagation” property

$k^\circ$ , the “oracle” choice:  $\max_{k \leq k^\circ} \Delta(W^{(k)}, \boldsymbol{\theta}) \leq \Delta$ .

▶ “Propagation” property: for  $k \leq k^\circ$

$$\mathbb{E} \log \left( 1 + \frac{|L(W^{(k)}, \tilde{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\theta}}_k)|^r}{\mathfrak{R}_r(\boldsymbol{\theta})} \right) \leq \Delta + \rho.$$

▶ Local constant GR:

$$\mathbb{E} \log \left( 1 + \frac{|(2\sigma^2)^{-1} N_k (\tilde{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_k)^2|^r}{\mathfrak{c}_r(\boldsymbol{\theta})} \right) \leq \Delta + \rho$$

▶ Local constant EF:

$$\mathbb{E} \log \left( 1 + \frac{|N_k \mathcal{K}(\tilde{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\theta}}_k)|^r}{\mathfrak{r}_r(\boldsymbol{\theta})} \right) \leq \Delta + \rho$$

## “Stability” properties

$k^\circ$ , the “oracle” choice:  $\max_{k \leq k^\circ} \Delta(W^{(k)}, \theta) \leq \Delta$ .

▶ “Stability” property:

$$L(W^{(k^\circ)}, \tilde{\theta}_{k^\circ}, \hat{\theta}) \mathbf{1}(\hat{k} \geq k^\circ) \leq 3k^\circ.$$

▶ Local constant GR:

$$(2\sigma^2)^{-1} N_{k^\circ} (\tilde{\theta}_{k^\circ} - \hat{\theta})^2 \mathbf{1}(\hat{k} \geq k^\circ) \leq 3k^\circ$$

▶ Local constant EF:

$$N_{k^\circ} \mathcal{K}(\tilde{\theta}_{k^\circ}, \hat{\theta}) \mathbf{1}(\hat{k} \geq k^\circ) \leq 3k^\circ$$

## “Oracle” result

Let  $\max_{k \leq k^\circ} \Delta(W^{(k)}, \theta) \leq \Delta$ . Then

$$\mathbb{E} \log \left( 1 + \frac{|L(W^{(k^\circ)}, \tilde{\theta}_{k^\circ}, \hat{\theta})|^r}{\mathfrak{R}_r(\theta)} \right) \leq \Delta + \rho + \log \left( \frac{\mathfrak{z}_{k^\circ}}{\mathfrak{R}_r(\theta)} \right).$$

► Local constant GR:

$$\mathbb{E} \log \left( 1 + \frac{|N_{k^\circ}(\tilde{\theta}_{k^\circ} - \hat{\theta})^2|^r}{(2\sigma^2)^r \mathfrak{c}_r} \right) \leq \Delta + \rho + \log \left( \frac{\mathfrak{z}_{k^\circ}}{\mathfrak{c}_r} \right).$$

► Local constant EF:

$$\mathbb{E} \log \left( 1 + \frac{|N_{k^\circ} \mathcal{K}(\tilde{\theta}_{k^\circ}, \hat{\theta})|^r}{\mathfrak{r}_r(\theta)} \right) \leq \Delta + \rho + \log \left( \frac{\mathfrak{z}_{k^\circ}}{\mathfrak{r}_r(\theta)} \right).$$

## Basic notions

---

Statistics is understanding data by modeling it.

Data  $\mathbf{Y} = (Y_1, \dots, Y_n)$  modeled as *random*.

$\mathbb{P} = \mathcal{L}(\mathbf{Y})$ , the *unknown* joint distribution.

Probabilistic problem: given  $\mathbb{P}$ , describe typical behavior of  $\mathbf{Y}$ .

**Statistical problem:** infer on  $\mathbb{P}$  from the data  $\mathbf{Y}$ .



## Parametric assumption (PA)

---

PA: the distribution  $\mathbb{P}$  of  $\mathbf{Y}$  is known up to the value of a  $p$ -dimensional parameter  $\boldsymbol{\theta}$ .

Equivalently  $\mathcal{L}(\mathbf{Y}) \in (\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p)$ .

$\boldsymbol{\theta}^*$  stands for the true parameter value:  $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*}$ .

Recovering  $\mathbb{P}$  is equivalent to estimating  $\boldsymbol{\theta}^*$  from  $\mathbf{Y}$ .

## Outline

---

### 1 Parameter Estimation. I.i.d. case

Estimation for i.i.d. sample: Examples  
ML estimation: Exponential family

### 2 Parametric Regression

(Mean) regression model  
Parametric estimation  
Estimation in regression-like model  
Estimation in Linear Gaussian model

### 3 ML and quasi ML estimation

Parameter estimation  
Exponential risk bound  
Pros and cont

## Empirical measure

Let  $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  where all  $Y_i$  are independent r.v.'s with distribution  $P$  on  $\mathbb{R}^1$ . Then  $\mathbb{P} = P^{\otimes n}$ .

Empirical measure  $P_n$ : for any measurable set  $A$

$$P_n(A) = \frac{1}{n} \sum \mathbf{1}(Y_i \in A).$$

### Theorem

Let  $g(\cdot)$  be a function on  $\mathbb{R}$  with

$$\int g(y) dP(y) = m, \quad \int [g(y) - m]^2 dP(y) = \sigma^2.$$

Then  $M_n \stackrel{\text{def}}{=} \int g(y) dP_n(y) = \frac{1}{n} \sum g(Y_i)$

satisfies  $\mathbb{E}M_n = m, \quad \text{Var}(M_n) = \sigma^2, \quad \sqrt{n}(M_n - m) \xrightarrow{w} \mathcal{N}(0, \sigma^2).$

## Substitution principle

---

**Idea:** express  $\theta^*$  as functional of  $P = P_{\theta^*}$  and use  $P_n$  instead of  $P$ .

**Substitution principle:** Let a functional  $G$  on the family  $(P_\theta)$  satisfy

$$G(P_\theta) \equiv \theta \quad \theta \in \Theta.$$

Then define

$$\tilde{\theta} \stackrel{\text{def}}{=} G(P_n)$$

**Method of moments. Univariate parameter**

I.i.d. sample  $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$  from  $P$ .

PA:  $P = P_{\theta^*} \in (P_\theta, \theta \in \Theta \subseteq \mathbb{R}^1)$ .

Let a function  $g(y)$  satisfy

$$\int g(y) dP_\theta(y) \equiv \theta,$$

$$\int [g(y) - \theta]^2 dP_\theta(y) = \sigma^2(\theta) < \infty.$$

Define

$$\tilde{\theta} \stackrel{\text{def}}{=} \int g(y) dP_n(y) = \frac{1}{n} \sum g(Y_i).$$

## Properties. Root-n consistency

### Theorem

The estimate  $\tilde{\theta} = n^{-1} \sum g(Y_i)$  fulfills

$$\mathbb{E}\tilde{\theta} = \theta^*,$$

$$\text{Var}(\tilde{\theta}) = \sigma^2(\theta^*)/n,$$

$$\sqrt{n}(\tilde{\theta} - \theta^*) \xrightarrow{w} \mathcal{N}(0, \sigma^2(\theta^*)).$$

## Properties. Concentration

### Theorem

Define for  $z > 0$

$$A(z) \stackrel{\text{def}}{=} \{\theta : |\theta - \theta^*| \leq zn^{-1/2}\sigma(\theta^*)\}.$$

Then

$$\mathbb{P}(\tilde{\theta} \notin A(z)) \rightarrow 2\Phi(-z)$$

**Interpretation:** The estimate  $\tilde{\theta}$  concentrates in a root-n interval around  $\theta^*$ .

## Properties. Confidence intervals

## Theorem

Define  $E(z) \stackrel{\text{def}}{=} \{\theta : |\theta - \tilde{\theta}| \leq zn^{-1/2}\sigma(\theta^*)\}, \quad z > 0.$

Then

$$\mathbb{P}(E(z) \not\ni \theta^*) \rightarrow 2\Phi(-z).$$

In particular, if  $z_\alpha$  satisfies  $2\Phi(-z_\alpha) = \alpha$ , then

$$\mathbb{P}(E(z_\alpha) \not\ni \theta^*) \rightarrow \alpha$$

**Interpretation:** The random interval  $E(z_\alpha)$  does not cover the true value  $\theta^*$  only with probability about  $\alpha$ .

Unfortunately, the construction of  $E(z_\alpha)$  depends upon  $\sigma(\theta^*)$  with  $\theta^*$  unknown.



**Properties. Confidence intervals. 2****Theorem**

Let  $\tilde{\sigma}$  be a consistent estimate of  $\sigma(\theta^*)$ . Define for  $z > 0$

$$\tilde{E}(z) \stackrel{\text{def}}{=} \{\theta : |\theta - \tilde{\theta}| \leq zn^{-1/2}\tilde{\sigma}\}.$$

Then

$$\mathbb{P}(\tilde{\theta} \notin \tilde{E}(z)) \rightarrow 2\Phi(-z)$$

In particular, if  $z_\alpha$  satisfies  $2\Phi(-z_\alpha) = \alpha$ , then

$$\mathbb{P}(\tilde{\theta} \notin \tilde{E}(z_\alpha)) \rightarrow \alpha.$$

## Extensions of MM

Let  $g(y)$  be a given function s.t. the function  $m(\theta)$  with

$$m(\theta) = \int g(y) dP_{\theta}(y)$$

is invertible. Then

$$\theta^* = m^{-1}(m(\theta^*)) = m^{-1}\left(\int g(y) dP_{\theta^*}(y)\right).$$

**MM approach:** Substitute  $P_{\theta^*}$  with its empirical counterpart  $P_n$ :

$$M_n \stackrel{\text{def}}{=} \int g(y) dP_n(y) = \frac{1}{n} \sum g(Y_i),$$

$$\tilde{\theta} = m^{-1}(M_n).$$

## Minimum distance estimates

---

Let  $\rho(P, Q)$  be a “distance” between two measures on  $\mathbb{R}^1$  s.t.

$$\rho(P, Q) \geq 0, \quad \rho(P, Q) = 0 \Leftrightarrow P = Q$$

Then

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} \rho(P_\theta, P_{\theta^*})$$

**Substitution:** replace  $P_{\theta^*}$  with  $P_n$ . Leads to the **Minimum Distance Estimate**

$$\tilde{\theta} \stackrel{\text{def}}{=} \operatorname{argmin}_{\theta \in \Theta} \rho(P_\theta, P_n).$$

## M-estimate

---

Let  $\psi(y, \theta)$  be a contrast function s.t.

$$\theta = \operatorname{argmin}_{\theta'} \int \psi(y, \theta') dP_{\theta}(y), \quad \theta \in \Theta.$$

In particular,

$$\theta^* = \operatorname{argmin}_{\theta \in \Theta} \int \psi(y, \theta) dP_{\theta^*}(y).$$

**Substitution:** replacing the true measure  $P_{\theta^*}$  with its empirical counterpart  $P_n$ :

$$\tilde{\theta} = \operatorname{argmin}_{\theta \in \Theta} \int \psi(y, \theta) dP_n(y) = \operatorname{argmin}_{\theta \in \Theta} \sum \psi(Y_i, \theta).$$

## Examples of M-estimates: Least Squares

The **least squares contrast** :  $\|\boldsymbol{\psi}(y) - \boldsymbol{\theta}\|^2$ , where  $\boldsymbol{\psi}(y)$  is a function of the observation  $y$  satisfying  $\int \boldsymbol{\psi}(y)dP_{\boldsymbol{\theta}}(y) \equiv \boldsymbol{\theta}$ . Then

$$\boldsymbol{\theta}^* = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \int \|\boldsymbol{\psi}(y) - \boldsymbol{\theta}\|^2 dP_{\boldsymbol{\theta}^*}(y).$$

and the M-estimation method leads to the **Least Squares Estimate (LSE)**:

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \int \|\boldsymbol{\psi}(y) - \boldsymbol{\theta}\|^2 dP_n(y) = \operatorname{argmin}_{\boldsymbol{\theta} \in \Theta} \sum \|\boldsymbol{\psi}(Y_i) - \boldsymbol{\theta}\|^2.$$

## Examples of M-estimates: Least Absolute Deviation

Consider the contrast  $\psi(y, \theta) = |y - \theta|$ .

### Lemma

For any measure  $P$  on  $\mathbb{R}$ , the median  $\text{med}(P)$  satisfies

$$\inf_{\theta \in \mathbb{R}} \int |y - \theta| dP(y) = \int |y - \text{med}(P)| dP(y).$$

If  $\theta \equiv \text{med}(P_\theta)$ , then

$$\theta = \underset{\theta'}{\text{argmin}} \int |y - \theta'| dP_\theta(y), \quad \theta \in \Theta.$$

Leads to the **Least Absolute Deviation** estimate

$$\tilde{\theta} \stackrel{\text{def}}{=} \underset{\theta \in \mathbb{R}}{\text{argmin}} \int |y - \theta| dP_n(y) = \underset{\theta \in \mathbb{R}}{\text{argmin}} \sum |Y_i - \theta|.$$

## Examples of M-estimates: Maximum Likelihood

Let  $\psi(y, \theta) = -\ell(y, \theta) = -\log p(y, \theta)$  where  $p(y, \theta)$  is the density of the measure  $P_\theta$  at  $y$  w.r.t. to some dominating measure  $\mu_0$ .

Leads to the **Maximum Likelihood Estimate (MLE)**:

$$\tilde{\theta} = \operatorname{argmax}_{\theta \in \Theta} n^{-1} \sum \log p(Y_i, \theta).$$

The condition on contrast is fulfilled because

$$\operatorname{argmin}_{\theta'} \int \log \frac{p(y, \theta)}{p(y, \theta')} dP_\theta(y) = \operatorname{argmin}_{\theta'} \mathcal{K}(\theta, \theta') = \theta,$$

where  $\mathcal{K}(\theta, \theta')$  is the Kullback-Leibler divergence for  $P_\theta$  and  $P_{\theta'}$ .

## Kullback-Leibler divergence

Kullback-Leibler (KL) divergence measures a “distance” between distributions  $P$  and  $Q$ :

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

In terms of parametric model  $P_{\theta}$ :

$$\mathcal{K}(\theta, \theta') = E_{\theta} \left\{ \log \left( \frac{dP_{\theta}}{dP_{\theta'}} \right) \right\}.$$

With pdf  $p(y, \theta)$ :

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



## Kullback-Leibler

---



Solomon Kullback (1903–1994), Richard A. Leibler (1914–2003)  
American mathematicians and cryptanalysts.

## Gaussian shift. Method of moments

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

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

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

It holds

$$\mathbb{E}_{\theta^*} Y_i = \theta^*, \quad \text{Var}(Y_i) = \sigma^2$$

Therefore MM-estimate is just the empirical mean:

$$\tilde{\theta} = n^{-1} \sum_{i=1}^n Y_i = \theta^* + \sigma n^{-1/2} \xi,$$

where  $\xi = \frac{1}{\sigma\sqrt{n}} \sum \varepsilon_i \sim \mathcal{N}(0, 1)$ .

## Gaussian shift: Confidence set

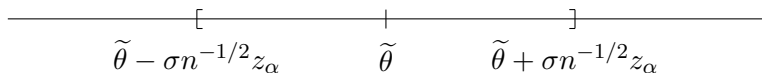
Let  $z_\alpha$  fulfill  $IP(|\xi| \leq z_\alpha) = 1 - \alpha$  for  $\xi \sim \mathcal{N}(0, 1)$ .

The decomposition  $\tilde{\theta} = \theta^* + \sigma n^{-1/2} \xi$  yields an  $\alpha$ -level confidence interval for  $\theta^*$

$$E(z_\alpha) = [\tilde{\theta} - \sigma n^{-1/2} z_\alpha, \tilde{\theta} + \sigma n^{-1/2} z_\alpha], \quad (1)$$

in the sense

$$\begin{aligned} IP_{\theta^*}(\mathcal{E}_\alpha \not\subseteq \theta^*) &= IP_{\theta^*}(|\tilde{\theta} - \theta^*| > \sigma n^{-1/2} z_\alpha) = \\ &= IP(|\xi| > z_\alpha) = \alpha. \end{aligned}$$



## Gaussian shift: ML approach

The log-likelihood for the Gaussian shift  $Y_i = \theta^* + \varepsilon_i$  reads as

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

Focus on  $\tilde{\theta} = \operatorname{argmax} L(\theta)$  and especially on the **maximum**

$$L(\tilde{\theta}) = \max_{\theta} L(\theta).$$

### Lemma

*It holds for any  $\theta$*

$$\begin{aligned} \tilde{\theta} &= n^{-1} S = n^{-1} \sum Y_i, \\ L(\tilde{\theta}, \theta) &\stackrel{\text{def}}{=} L(\tilde{\theta}) - L(\theta) = n\sigma^{-2}(\tilde{\theta} - \theta)^2/2. \end{aligned}$$

## Gaussian shift: Wilks phenomenon

The decomposition  $\tilde{\theta} = \theta^* + \sigma n^{-1/2} \xi$  implies

### Theorem

*It holds*

$$2L(\tilde{\theta}, \theta^*) = n\sigma^{-2}(\tilde{\theta} - \theta^*)^2 = \xi^2 \sim \chi_1^2.$$

If  $z_{\alpha}$  is the  $\alpha$ -quantile of  $\chi_1^2$  with  $P(\xi^2 > z_{\alpha}) = \alpha$ , then

$$\mathcal{E}(z_{\alpha}) = \{u : L(\tilde{\theta}, u) \leq z_{\alpha}\}$$

is again an  $\alpha$ -CS (actually the same as  $E(z_{\alpha})$ ), but this time “likelihood based”.

## Bernoulli model

---

Let  $Y_1, \dots, Y_n$  be i.i.d. Bernoulli r.v.'s satisfying

$$P_{\theta}(Y_i = 1) = \theta, \quad P_{\theta}(Y_i = 0) = 1 - \theta.$$

Examples:

- ▶ Coin throws
- ▶ Binary signals and images
- ▶ Binary choice models

**Bernoulli model: MM-estimation**

Observe

$$E_{\theta} Y_i = \theta, \quad E_{\theta} (Y_i - \theta)^2 = \theta(1 - \theta).$$

Yields the **MM-estimate**

$$\tilde{\theta} = \frac{1}{n} \sum Y_i$$

and **asymptotic confidence sets**

$$\tilde{E}(z_{\alpha}) = \{\theta : |\theta - \tilde{\theta}| \leq z_{\alpha} n^{-1/2} \tilde{\sigma}\}$$

where  $\tilde{\sigma}^2$  is an estimate of  $\sigma^2(\theta^*) \stackrel{\text{def}}{=} \theta^*(1 - \theta^*)$ .

## Bernoulli model: ML Estimation

It holds

$$L(\theta) = \log \prod_{i=1}^n \theta^{Y_i} (1 - \theta)^{1 - Y_i} = \log \theta \sum_i Y_i + \log(1 - \theta) \sum_i (1 - Y_i)$$

### Lemma

For any  $\theta$

$$\tilde{\theta} = S/n = n^{-1} \sum Y_i \quad L(\tilde{\theta}, \theta) = n\mathcal{K}(\tilde{\theta}, \theta)$$

where  $S = Y_1 + \dots + Y_n$  and

$$\mathcal{K}(\theta, \theta') = \theta \log(\theta/\theta') + (1 - \theta) \log[(1 - \theta)(1 - \theta)']$$

is the *Kullback-Leibler divergence* for the Bernoulli law.



## Poisson model

---

Let  $Y_1, \dots, Y_n$  be i.i.d. Poisson r.v.'s satisfying

$$\mathbb{P}_\theta(Y_i = m) = \theta^m e^{-\theta} / m! \quad m = 0, 1, 2, \dots$$

### Examples:

- ▶ Number of telephone calls arriving at a switchboard / an automatic phone-switching system
- ▶ Number of web page requests arriving at a server except for unusual circumstances such as coordinated denial of service attacks
- ▶ number of photons registered in a cell for digital imaging or PET
- ▶ number of stars observed in a sky segment

## Poisson model: MM-estimation

---

Observe

$$E_{\theta} Y_i = \theta, \quad E_{\theta} (Y_i - \theta)^2 = \theta.$$

Yields the **MM-estimate**

$$\tilde{\theta} = \frac{1}{n} \sum Y_i$$

and **asymptotic confidence sets**

$$\tilde{E}(z_{\alpha}) = \{\theta : |\theta - \tilde{\theta}| \leq z_{\alpha} n^{-1/2} \tilde{\sigma}\}$$

where  $\tilde{\sigma}^2$  is an estimate of  $\sigma^2(\theta^*) \stackrel{\text{def}}{=} \theta^*$ .

## Poisson model: ML Estimation

For  $Y_i$  i.i.d. from  $\text{Poisson}(\theta)$

$$L(\theta) = \log \prod_{i=1}^n \theta^{Y_i} e^{-\theta} / Y_i! = \log \theta \sum_{i=1}^n Y_i - n\theta - \sum_{i=1}^n \log(Y_i!)$$

### Lemma

For any  $\theta$

$$\tilde{\theta} = S/n = n^{-1} \sum Y_i \quad L(\tilde{\theta}, \theta) = n\mathcal{K}(\tilde{\theta}, \theta)$$

where  $S = Y_1 + \dots + Y_n$  and  $\mathcal{K}(\theta, \theta') = \theta \log(\theta/\theta') - (\theta - \theta')$  is the Kullback-Leibler divergence for the Poisson law.

## Poisson model: Details

---

$$L(\theta, \theta') = S \log(\theta/\theta') - n(\theta - \theta'),$$

where  $S = Y_1 + \dots + Y_n$  and

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

while

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

## Exponential model

---

Let  $Y_1, \dots, Y_n$  be i.i.d. exponential r.v.'s with parameter  $\theta > 0$ :

$$P_{\theta}(Y_i > t) = e^{-t/\theta}.$$

### Examples:

- ▷ Intervals between transactions
- ▷ Waiting time in a queue
- ▷ Time to a failure

## Exponential model: MM Estimation

Observe

$$E_{\theta} Y_i = \theta, \quad E_{\theta} (Y_i - \theta)^2 = \theta^2.$$

Yields the **MM-estimate**

$$\tilde{\theta} = \frac{1}{n} \sum Y_i$$

and **asymptotic confidence sets**

$$\tilde{E}(z_{\alpha}) = \{\theta : |\theta - \tilde{\theta}| \leq z_{\alpha} n^{-1/2} \tilde{\sigma}\}$$

where  $\tilde{\sigma}^2$  is an estimate of  $\sigma^2(\theta^*) \stackrel{\text{def}}{=} \theta^{*2}$ .

## Exponential model: ML Estimation

With  $\ell(y, \theta) = -\log \theta - y/\theta$

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

### Lemma

For any  $\theta$

$$\tilde{\theta} = S/n = n^{-1} \sum Y_i \quad L(\tilde{\theta}, \theta) = n\mathcal{K}(\tilde{\theta}, \theta)$$

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

## Volatility model

---

Let  $\xi_1, \dots, \xi_n$  be i.i.d.  $\mathcal{N}(0, \theta)$  r.v.'s. Observed  $Y_i = \xi_i^2$ .

### Examples:

- ▶ squared log-returns of a stock.
- ▶ Errors in regression.



## Volatility model: MM Estimation

Observe

$$E_{\theta} Y_i = \theta, \quad E_{\theta} (Y_i - \theta)^2 = 2\theta^2.$$

Yields the **MM-estimate**

$$\tilde{\theta} = \frac{1}{n} \sum Y_i$$

and **asymptotic confidence sets**

$$\tilde{E}(z_{\alpha}) = \{\theta : |\theta - \tilde{\theta}| \leq z_{\alpha} n^{-1/2} \tilde{\sigma}\}$$

where  $\tilde{\sigma}^2$  is an estimate of  $\sigma^2(\theta^*) \stackrel{\text{def}}{=} 2\theta^{*2}$ .

## Volatility model: ML Estimation

With  $\ell(y, \theta) = -1/2 \log(2\pi\theta) - y/(2\theta)$

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

### Lemma

For any  $\theta$

$$\tilde{\theta} = S/n = n^{-1} \sum Y_i \quad L(\tilde{\theta}, \theta) = n\mathcal{K}(\tilde{\theta}, \theta)$$

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

## Exponential family (EF)

---

In an **exponential family** (EF), all measures  $P_\theta$  have pdfs:

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

$$\ell(y, \theta) = yC(\theta) - B(\theta) + \log p(y).$$

Here

- ▶  $C(\theta)$  and  $B(\theta)$  - monotonous functions of  $\theta$
- ▶  $p(y)$  - nonnegative function.

## Exponential family. Natural parametrization

The *natural parametrization* means the relation

$$E_{\theta}Y = \theta$$

### Lemma

Let  $\mathcal{P} = (P_{\theta})$  be an EF with natural parametrization (EFn). Then

- ▷  $B'(\theta) = \theta C'(\theta)$  ;
- ▷  $\text{Var}_{\theta}(Y) = 1/C'(\theta)$  ;
- ▷ the KL divergence  $\mathcal{K}(\theta, \theta') \stackrel{\text{def}}{=} E_{\theta} \log \left\{ \frac{p(Y, \theta)}{p(Y, \theta')} \right\}$  satisfies

$$\mathcal{K}(\theta, \theta') = \theta \{C(\theta) - C(\theta')\} - \{B(\theta) - B(\theta')\}, \quad \theta, \theta' \in \Theta;$$

- ▷ For Fisher information, it holds:  $I(\theta) \stackrel{\text{def}}{=} E_{\theta} \left| \frac{\partial}{\partial \theta} \ell(y, \theta) \right|^2 = C'(\theta)$ .

## KL divergence for some EF's

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

Model	$\mathcal{K}(\theta, \theta')$
Gaussian	$(\theta - \theta')^2 / (2\sigma^2)$
Bernoulli	$\theta \log(\theta/\theta') + (1 - \theta) \log\{(1 - \theta)/(1 - \theta')\}$
Poisson	$\theta \log(\theta/\theta') - (\theta - \theta')$
Exponential	$\theta/\theta' - 1 - \log(\theta/\theta')$
Volatility	$\frac{1}{2}(\theta/\theta' - 1) - \frac{1}{2} \log(\theta/\theta')$

## Fisher information for some EF's

$$I(\theta) = \mathbb{E} \left\{ \frac{\partial}{\partial \theta} \ell(Y, \theta) \right\}^2 = C'(\theta).$$

Model	$I(\theta)$
Gaussian	$\sigma^{-2}$
Bernoulli	$\theta^{-1}(1 - \theta)^{-1}$
Poisson	$\theta^{-1}$
Exponential	$\theta^{-2}$
Volatility	$(2\theta^2)^{-1}$

## Exponential family. Canonical parametrization

The *canonical* parametrization means that  $\ell(y, \theta)$  is linear w.r.t.  $\theta$ :

$$\ell(y, \theta) = y\theta - d(\theta)$$

where  $d(\cdot)$  is a convex function.

### Lemma

Let  $\mathcal{P} = (P_\theta)$  be an EF with canonical parametrization (EFC). Then

- ▶  $E_\theta Y = d'(\theta)$  and  $\text{Var}_\theta Y = I_\theta = d''(\theta)$ ;
- ▶ the KL divergence  $\mathcal{K}(\theta, \theta') = E_\theta \log \left\{ \frac{p(Y, \theta)}{p(Y, \theta')} \right\}$  satisfies

$$\mathcal{K}(\theta, \theta') = d(\theta') - d(\theta) - (\theta' - \theta)d'(\theta), \quad \theta, \theta' \in \Theta.$$

- ▶ For Fisher information, it holds:  $I(\theta) \stackrel{\text{def}}{=} E_\theta \left| \frac{\partial}{\partial \theta} \ell(y, \theta) \right|^2 = d''(\theta)$ .

## MM-estimation for EFn

The relations

$$E_{\theta} Y_i = \theta, \quad \text{Var}_{\theta} Y_i = 1/C'(\theta)$$

yield the **MM-estimate**

$$\tilde{\theta} = \frac{1}{n} \sum Y_i$$

and the **asymptotic confidence sets**

$$\tilde{E}(z_{\alpha}) = \{\theta : |\tilde{\theta} - \theta| \leq z_{\alpha} n^{-1/2} \tilde{\sigma}\}$$

where  $\tilde{\sigma}^2$  estimates  $1/C'(\theta^*)$ .



## ML-approach for EFn

The log-density  $\log p(y, \theta) = yC(\theta) - B(\theta) + \log p(y)$  leads to the log-likelihood

$$L(\theta) = \sum_{i=1}^n \log p(Y_i, \theta) = S C(\theta) - nB(\theta) + R$$

where  $S = \sum_{i=1}^n Y_i$  and  $R = \sum \log p(Y_i)$ .

### Lemma

For any  $\theta$

$$\tilde{\theta} = S/n = n^{-1} \sum Y_i \quad L(\tilde{\theta}, \theta) = n\mathcal{K}(\tilde{\theta}, \theta)$$

## Exponential bounds for the fitted likelihood. EF case

Maximum likelihood:

$$L(\tilde{\theta}, \theta^*) \stackrel{\text{def}}{=} \max_{\theta} \{L(\theta) - L(\theta^*)\}.$$

## Theorem (Polzehl and Spokoiny (2005))

Let  $(P_{\theta})$  be an EF. Then for any  $\mathfrak{z} > 0$  and  $r > 0$

$$\mathbb{P}_{\theta^*} \{L(\tilde{\theta}, \theta^*) > \mathfrak{z}\} = \mathbb{P}_{\theta^*} \{n\mathcal{K}(\tilde{\theta}, \theta^*) > \mathfrak{z}\} \leq 2e^{-\mathfrak{z}},$$

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

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

**Interpretation:**  $L(\tilde{\theta}, \theta^*) = n\mathcal{K}(\tilde{\theta}, \theta^*)$  is stochastically bounded whatever EF and sample size  $n$  are.

## Discussion

- ▶  $\theta^* = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_{\theta^*} L(\theta, \theta^*) = -n\mathcal{K}(\theta^*, \theta^*) .$
- ▶  $\tilde{\theta} = \operatorname{argmax}_{\theta \in \Theta} L(\theta, \theta^*) .$
- ▶ By PS2005  $\tilde{\theta}$  is close to  $\theta^*$  in the sense that  $L(\tilde{\theta}) - L(\theta^*) = n\mathcal{K}(\tilde{\theta}, \theta^*)$  is stochastically bounded.

### Corollary (Likelihood-based confidence sets)

Define  $\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \{\theta : L(\tilde{\theta}, \theta) \leq \mathfrak{z}\} = \{\theta : n\mathcal{K}(\tilde{\theta}, \theta)\} .$  Then

$$\mathbb{P}_{\theta^*}(\mathcal{E}(\mathfrak{z}) \not\ni \theta^*) \leq 2e^{-\mathfrak{z}} .$$

## Outline

---

### 1 Parameter Estimation. I.i.d. case

Estimation for i.i.d. sample: Examples  
ML estimation: Exponential family

### 2 Parametric Regression

(Mean) regression model  
Parametric estimation  
Estimation in regression-like model  
Estimation in Linear Gaussian model

### 3 ML and quasi ML estimation

Parameter estimation  
Exponential risk bound  
Pros and cont

## Regression model

---

The (mean) regression model links the *explained variable*  $Y$  and the *explanatory variable*  $X$  in the form

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

Equivalent formulation:

$$E(Y|X = x) = f(x).$$

## Regression model: Ingredients

---

- ▷ *Observations*  $(X_i, Y_i)$  for  $i = 1, \dots, n$ .  $n$  is the *sample size*.
  - ▷  $Y_i$  are independent.
  - ▷  $Y_i$  progressively dependent (time series);
  - ▷  $Y_i$  mutually dependent;
  
- ▷ *Design*  $X_1, \dots, X_n$ ,  $X_i \in \mathcal{X}$  where  $\mathcal{X}$  is the design space.
  - ▷ **Deterministic**;
  - ▷ Random with a density  $p(x)$ ;
  - ▷ Continuous/discrete/mixed;

## Regression model: Ingredients

---

- ▶ Errors  $\varepsilon_i$ . In general zero mean:

$$\mathbb{E}(\varepsilon|X) = 0.$$

Typical assumption:

- ▶ *Homoscedastic errors*:  $\text{Var } \varepsilon_i = \sigma^2$ .
- ▶ *Heteroscedastic errors*:  $\text{Var } \varepsilon_i$  depends on the location  $X_i$  or the value  $f(X_i)$ .
- ▶ *Regression function*  $f(x)$  for  $x \in \mathcal{X}$ .
  - ▶ The parametric case:  $f(x) = f(x, \theta)$  is known up to a parameter  $\theta \in \Theta \subset \mathbb{R}^p$ .
  - ▶ *Nonparametric case*:  $f(x)$  is smooth in the sense that it admits a good local polynomial approximation.

**Example: Wage equation**

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

where  $Y = \log \text{wages}$ ,  $X = (X_1, X_2, X_3)^\top$  with

1.  $X_1 = \text{schooling}$  (measured in years);
2.  $X_2 = \text{labor market experience}$   
(measured as:  $\text{AGE} - \text{SCHOOL} - 6$ );
3.  $X_3 = \text{experience squared}$ .

CPS 1985,  $n = 534$ , see Berndt (1991).

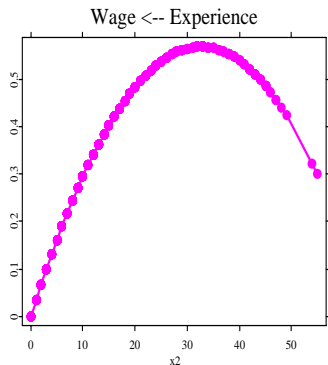
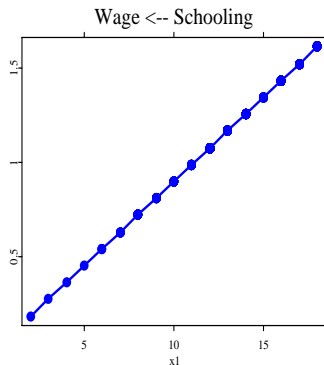


**Coefficient estimates for the wage equation:**

Dependent Variable: Log Wages			
Variable	Coefficients	S.E.	<i>t</i> -values
SCHOOL	0.0898	0.0083	10.788
EXP	0.0349	0.0056	6.185
EXP <sup>2</sup>	-0.0005	0.0001	-4.307
constant	0.5202	0.1236	4.209

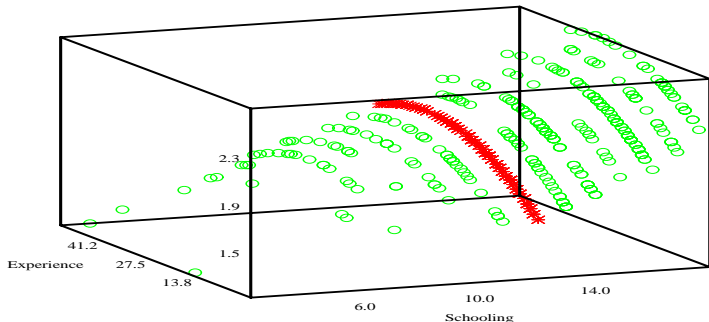
$R^2 = 0.24$ , sample size  $n = 534$

**Table:** Results from ordinary LS estimation MNScps85lin



wage-schooling profile and wage-experience profile MNScsp85lin

Wage  $\leftarrow$  Schooling, Experience



Parametrically estimated regression function `MNScps85lin`

## Nonparametric Regression

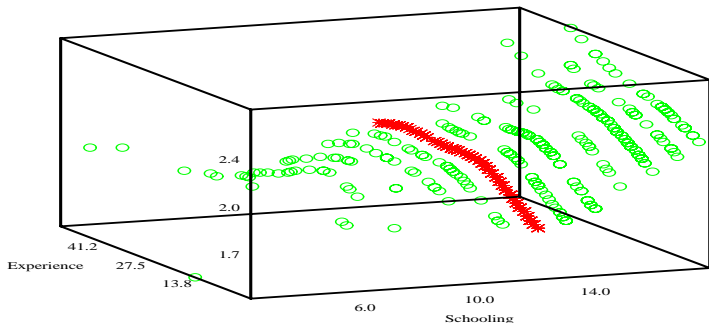
---

With  $X = (X_1, X_2)^\top = (\text{SCHOOL}, \text{EXP})^\top$

$$E(Y|X = x) = f(x)$$

where  $f(\cdot)$  is a smooth function.

Wage  $\leftarrow$  Schooling, Experience



Nonparametrically estimated regression function MNScsp85reg

## Parametric regression. Substitution

Regression model:

$$Y_i = f(X_i) + \varepsilon_i \quad \varepsilon_i \text{ i.i.d. } p(\cdot)$$

Target: regression function  $f$ .

$$PA : f(\cdot) = f(\cdot, \theta^*)$$

$f(\cdot)$  is known up to a finite dimensional parameter  $\theta^* \in \Theta \subseteq \mathbb{R}^p$ .  
Can be rewritten in terms of **residuals**:

$$\varepsilon_i = Y_i - f(X_i, \theta^*).$$

**Substitution approach**: select  $\theta^*$  to provide the best fit of the empirical distribution of residuals to their population counterpart.

## Parametric M-estimation

Let  $\psi(z)$  be a contrast function s.t. for any  $z$

$$\mathbb{E}\psi(z + \varepsilon_i) \geq \mathbb{E}\psi(z)$$

*M-estimate:*

$$\tilde{\theta} = \underset{\theta}{\operatorname{argmin}} \sum \psi \left\{ Y_i - f(X_i, \theta) \right\}.$$

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

## Examples: MLE for Linear regression

---

Let  $\psi_1(x), \dots, \psi_p(x)$  be given basis functions and

$$f(x, \boldsymbol{\theta}) = \theta_1 \psi_1(x) + \dots + \theta_p \psi_p(x)$$

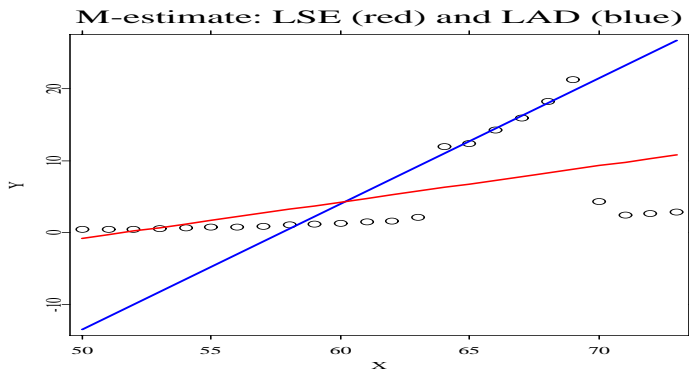
Then

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \sum \ell(Y_i - \boldsymbol{\theta}^\top \boldsymbol{\Psi}_i),$$

where  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$  and  $\boldsymbol{\Psi}_i = (\psi_1(X_i), \dots, \psi_p(X_i))^\top$ .



## Example. International phone calls from Belgium



Liner Regression  $f(x, \theta) = \theta_0 + \theta_1 x$  based on the international phone calls from Belgium in years 1950-1973.

## Regression-like model

---

Let  $\mathcal{P} = (P_\nu, \nu \in \mathcal{U} \subseteq \mathbb{R})$  be a parametric family, dominated by  $P$ , and  $p(y, \nu) = dP_\nu/dP(y)$ .

*Regression-like model:*  $Y_i$  are independent and the distribution of  $Y_i$  belongs to  $\mathcal{P}$  where the parameter depends on  $X_i$  through  $f(X_i)$ :

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

The *regression function*  $f(\cdot)$  identifies the distribution of  $\mathbf{Y}$ :

$$L(f) = \sum_{i=1}^n \log p(Y_i, f(X_i)).$$

## Varying coefficient EF modeling

---

Model:

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

In standard cases  $\mathcal{P}$  is an exponential family (EF) with the **natural** ( $E_v Y = v$ ) or **canonical** ( $\log p(y, v)$  is linear in  $v$ ) parametrization.

For the natural parametrization

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

Referred to as *Varying coefficient (nonparametrically driven) exponential family*.

## Regression-like parametric models

---

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

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n \ell\{Y_i, f(X_i, \boldsymbol{\theta})\}$$

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

**Example: Constant regression for an EFn**

Let  $\mathcal{P} = (P_v)$  be an EF with the natural parametrization:

$$E_v Y = v.$$

Let  $\theta \in \mathcal{U}$  and  $f(x, \theta) = \theta$ . Then

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

## Generalized Linear regression

Model  $Y_i \sim P_{f(X_i)} \in \mathcal{P}$  where  $\mathcal{P}$ , an EF with **canonical** parametrization with  $\ell(y, v) = yv - d(v) + \log p(y)$  and  $E_v Y = d'(v)$ .

**Generalized linear modeling:**  $f(X_i) = \boldsymbol{\theta}^\top \boldsymbol{\Psi}_i$  where  $\boldsymbol{\Psi}_i = \boldsymbol{\Psi}(X_i)$  is a given vector of features.

Leads to the MLE

$$\tilde{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_i \{Y_i \boldsymbol{\theta}^\top \boldsymbol{\Psi}_i - d(\boldsymbol{\theta}^\top \boldsymbol{\Psi}_i)\}.$$

This is a convex optimization problem but in general no close form solution.

## GL Modeling

Estimating equation:  $\nabla L(\tilde{\theta}) = \sum_i Y_i \Psi_i - \sum_i \Psi_i d'(\tilde{\theta}^\top \Psi_i) = 0$ .

Leads to the representation

$$\begin{aligned}\nabla L(\tilde{\theta}) - \nabla L(\theta^*) &= B(\theta') (\tilde{\theta} - \theta^*), \\ \tilde{\theta} - \theta^* &= -B^{-1}(\theta') \sum_i \{Y_i - d'(\theta^*)\} \Psi_i\end{aligned}$$

where  $\theta'$  is on the line between  $\theta$  and  $\tilde{\theta}$  and

$$B(\theta) = \nabla^2 L(\theta) = \sum_i \Psi_i \Psi_i^\top d''(\theta^\top \Psi_i).$$

Quadratic expansion of  $L(\tilde{\theta}, \theta)$  at  $\tilde{\theta}$ : for any  $\theta$  and some  $\theta^\circ \in [\theta, \tilde{\theta}]$

$$L(\tilde{\theta}, \theta) = 0.5(\tilde{\theta} - \theta)^\top B(\theta^\circ)(\tilde{\theta} - \theta)$$

## Example. Eastern Western German Immigration





## Example continued

$$Y = \begin{cases} 1 & \text{if person imagines to move to west,} \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}(Y|X) = P(Y = 1|X) = G(\beta^\top X),$$

where  $X$ , a vector of personal features, and  $f(X) = G(\beta^\top X)$  the related parameter.

Leads to the log-likelihood

$$L(\beta) = \sum_{i=1}^n \left[ Y_i \log \frac{G(\beta^\top X_i)}{1 - G(\beta^\top X_i)} + \log \{1 - G(\beta^\top X_i)\} \right].$$

## Example continued

---

The choice of the logistic link function  $G(u) = (1 + e^{-u})^{-1}$  (logit model) corresponds to the canonical parametrization:

$$L(\beta) = \sum_{i=1}^n \{Y_i \beta^\top X_i + \log(1 + e^{\beta^\top X_i})\}.$$

## Example continued

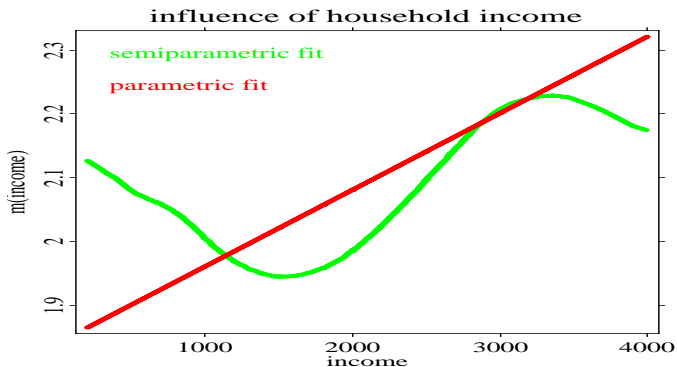
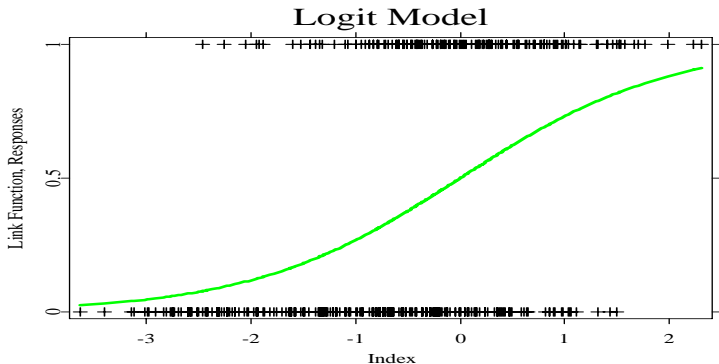


Figure: Estimated influence of income  $\tilde{f}(t)$

## Example continued



Logit model for migration MNSlogit



## Summary

(Mean) regression model:

$$Y_i = f(X_i) + \varepsilon_i$$

Regression-like model

$$Y_i \sim P_{f(X_i)}.$$

Parametric ML-estimation:  $f(x) = f(x, \boldsymbol{\theta})$

$$\tilde{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^n \ell\{Y_i, f(X_i, \boldsymbol{\theta})\}.$$

## Linear Model

Consider the model

$$Y_i = \Psi_i^\top \boldsymbol{\theta}^* + \varepsilon_i, \quad n = 1, \dots, n,$$

- ▶  $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_p^*)^\top \in \mathbb{R}^p$ , an unknown parameter vector,
- ▶  $\Psi_i$ , given vectors in  $\mathbb{R}^p$  and
- ▶  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^\top$ , Gaussian errors with zero mean and a **known** covariance matrix  $\Sigma$ :  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ .

Special cases:

1.  $\varepsilon_i$  are i.i.d.  $\mathcal{N}(0, \sigma^2)$ , or equivalently,  $\Sigma = \sigma^2 I_n$ .
2.  $\varepsilon_i$  are independent,  $\mathbb{E}\varepsilon_i^2 = \sigma_i^2$ . Then  $\Sigma$  is diagonal:  
 $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$ .

## ML-approach

The model equation can be rewritten in vector form:

$$\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma).$$

yielding the log-likelihood

$$L(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) - \frac{\log(\det \Sigma)}{2} - \frac{1}{2} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1} (\mathbf{Y} - \Psi^\top \boldsymbol{\theta}).$$

In case 1 this expression can be rewritten as

$$L(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (Y_i - \Psi_i \boldsymbol{\theta})^2.$$

In case 2 the expression is similar:

$$L(\boldsymbol{\theta}) = - \sum \left\{ \frac{1}{2} \log(2\pi\sigma_i^2) + \frac{(Y_i - \Psi_i \boldsymbol{\theta})^2}{2\sigma_i^2} \right\}.$$

## MLE

The Maximum Likelihood Estimate (MLE)  $\tilde{\theta}$  of  $\theta^*$  is defined by maximizing the log-likelihood  $L(\theta)$ :

$$\tilde{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmax}} L(\theta) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} (\mathbf{Y} - \Psi^\top \theta)^\top \Sigma^{-1} (\mathbf{Y} - \Psi^\top \theta). \quad (2)$$

Differentiating the right hand-side of (2) w.r.t.  $\theta$  yields the *normal equation*

$$\Psi \Sigma^{-1} \Psi^\top \tilde{\theta} = \Psi \Sigma^{-1} \mathbf{Y}.$$

If the  $p \times p$ -matrix  $\Psi \Sigma^{-1} \Psi^\top$  is non degenerated then

$$\tilde{\theta} = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1} \mathbf{Y} = \Phi \mathbf{Y},$$

where  $\Phi = (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$  is a fixed  $p \times n$  matrix.



## Response estimation and maximum likelihood

The vector  $\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}}$  is an estimate of the response  $\mathbf{f} \stackrel{\text{def}}{=} \mathbb{E}\mathbf{Y}$  :

$$\tilde{\mathbf{f}} = \Psi^\top \tilde{\boldsymbol{\theta}} = \Psi^\top (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1} \mathbf{Y} = \Pi \mathbf{Y},$$

where  $\Pi = \Psi^\top (\Psi \Sigma^{-1} \Psi^\top)^{-1} \Psi \Sigma^{-1}$  is a  $n \times n$  matrix (linear operator).

### Theorem

For any  $\boldsymbol{\theta}$  holds

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \Psi \Sigma^{-1} \Psi^\top (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \|\Sigma^{-1/2}(\tilde{\mathbf{f}} - \mathbf{f}_\boldsymbol{\theta})\|^2$$

where  $\mathbf{f}_\boldsymbol{\theta} = \Psi^\top \boldsymbol{\theta}$ . In particular, if  $\Sigma = \sigma^2 I_n$  then the fitted log-likelihood is proportional to the quadratic loss  $\|\tilde{\mathbf{f}} - \mathbf{f}_\boldsymbol{\theta}\|^2$  :

$$2\sigma^2 L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \|\Psi^\top (\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 = \|\tilde{\mathbf{f}} - \mathbf{f}_\boldsymbol{\theta}\|^2.$$

## Wilks phenomenon and Confidence Ellipsoid

### Theorem (Wilks phenomenon)

Assume  $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$  with  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ . Then

$$2L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \sim \chi_p^2 \quad \text{chi-squared with } p \text{ degrees of freedom}$$

This result can be used to build confidence ellipsoids for  $\boldsymbol{\theta}^*$ .

### Theorem (Confidence ellipsoids)

Assume  $\mathbf{Y} = \Psi^\top \boldsymbol{\theta}^* + \boldsymbol{\varepsilon}$  with  $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Sigma)$ . Define  $\mathfrak{z}_\alpha$  by  $P\{\chi_p^2 > 2\mathfrak{z}_\alpha\} = \alpha$ . Then

$$\mathcal{E}(\mathfrak{z}_\alpha) = \{\boldsymbol{\theta} : L(\tilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) \leq \mathfrak{z}_\alpha\}$$

is an  $\alpha$ -level confidence set for  $\boldsymbol{\theta}^*$ .

## Outline

---

### 1 Parameter Estimation. I.i.d. case

Estimation for i.i.d. sample: Examples

ML estimation: Exponential family

### 2 Parametric Regression

(Mean) regression model

Parametric estimation

Estimation in regression-like model

Estimation in Linear Gaussian model

### 3 ML and quasi ML estimation

Parameter estimation

Exponential risk bound

Pros and cont

## Parametric model. Likelihood

Data  $\mathbf{Y}$ .  $P = \mathcal{L}(\mathbf{Y})$ .

PA:  $P \in (P_\theta, \theta \in \Theta \subseteq \mathbb{R}^p)$  or  $P = P_{\theta^*}$  for some  $\theta^* \in \Theta$ .

Let  $P_\theta \ll \mu_0$  for some measure  $P$  for all  $\theta \in \Theta$ . Define the **log-likelihood**

$$L(\theta) = \log \frac{dP_\theta}{d\mu_0}(\mathbf{Y}).$$

For some  $\theta^\circ \in \Theta$ , **the log-likelihood ratio** is

$$L(\theta, \theta^\circ) = L(\theta) - L(\theta^\circ) = \log \frac{dP_\theta}{dP_{\theta^\circ}}(\mathbf{Y}).$$

## General Maximum Likelihood (ML) approach

---

- ▶ *Maximum likelihood estimate*  $\tilde{\theta}$  is the point of maximum of  $L(\theta)$ :

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

- ▶ Focus on **Maximum Likelihood**:

$$L(\tilde{\theta}, \theta^\circ) = \max_{\theta \in \Theta} L(\theta, \theta^\circ).$$

- ▶ The quality of estimation is measured by  $L(\tilde{\theta}, \theta^*) = L(\tilde{\theta}) - L(\theta^*)$  rather than by  $\tilde{\theta} - \theta^*$ .

## Quasi ML approach

Let  $P = \mathcal{L}(Y)$  and let  $(P_\theta)$  be a given parametric family.

► The PA “ $P \in (P_\theta)$ ” is **possibly wrong**, however one proceeds as if it is fulfilled. Leads to the value

$$\tilde{\theta} = \operatorname{argmax}_{\theta \in \Theta} L(\theta).$$

► The **target** of estimation  $\theta^*$  is defined as

$$\theta^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}L(\theta)$$

$\theta^*$  defines the **best parametric fit** of  $P$  by  $(P_\theta)$ .

►  $\tilde{\theta}$  is again an **empirical counterpart (estimate)** of  $\theta^*$ .

## Examples

Parametric regression model  $\mathbb{E}(Y_i|X_i) = f(X_i, \boldsymbol{\theta})$ .

Least Squares Estimate (LSE):

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

Becomes MLE if  $\varepsilon_i$  are  $\mathcal{N}(0, \sigma^2)$ , otherwise quasi MLE.

Least Absolute Deviation (LAD):

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

Becomes MLE if the  $\varepsilon_i$ 's are Laplacian (double exponential).

## A general exponential bound

Data:  $Y \sim IP$ . PA:  $IP \in (IP_{\theta}, \theta \in \Theta \subseteq \mathbb{R}^p)$ . Possibly **wrong**.

best parametric fit:  $\theta^* \stackrel{\text{def}}{=} \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}L(\theta)$ .

### Theorem (Golubev and Spokoiny (2009))

*Under regularity conditions, for  $\mu \in (0, 1)$*

$$\mathbb{E}_{\theta^*} \exp\{\mu L(\tilde{\theta}, \theta^*)\} \leq \mathfrak{Q}(\mu, \theta^*) \leq \mathfrak{Q}(\mu, \Theta).$$

*where  $\mathfrak{Q}(\Theta)$  is some fixed constant.*



## Some corollaries. Likelihood-based confidence sets

---

The exponential bound on  $L(\tilde{\theta}, \theta^*)$  implies **confidence sets** of the form

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

Indeed,

$$P_{\theta^*}(\mathcal{E}(\mathfrak{z}) \not\ni \theta^*) = P_{\theta^*}(L(\tilde{\theta}, \theta^*) > \mathfrak{z}) \leq e^{-\mu\mathfrak{z}}\Omega(\mu, \Theta) \rightarrow 0, \quad \mathfrak{z} \rightarrow \infty.$$

However, the exponential bound of Theorem 21 is not sharp. A careful choice of the parameter  $\mathfrak{z}$  that ensures the prescribed level should be done by some resampling methods.

## Some corollaries. Root-n consistence

Exponential bound implies for any  $r > 0$

$$\mathbb{E}_{\theta^*} |L(\tilde{\theta}, \theta^*)|^r \leq \mathfrak{R}_r(\theta^*) \leq \mathfrak{R}_r(\theta),$$

where  $\mathfrak{R}_r(\theta)$  is some fixed constant.

In regular cases:

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

where  $I(\theta^*)$ , the Fisher information matrix. Theorem implies

$$\mathbb{E}_{\theta^*} |\sqrt{n/2} \sqrt{I(\theta^*)}(\tilde{\theta} - \theta^*)|^{2r} \leq \mathfrak{R}_r(\theta^*),$$

and yields **root-n consistency**:

$$\mathbb{E}_{\theta^*}^{1/(2r)} |\sqrt{I(\theta^*)}(\tilde{\theta} - \theta^*)|^{2r} \leq c/\sqrt{n}$$

## Parametric modeling. Pros

---

1. Well developed algorithms
2. Nice nonasymptotic theory. Implies risk bounds and exact confidence sets.
3. Good in-sample properties.

## Drawbacks of parametric modeling

---

The parametric structure is crucial. If the parametric assumption is violated, the MLE estimator  $\tilde{\theta}$  is misspecified.

A parametric model may be like a Procrustes bet for the data: “cut off” of important features.

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