Weierstraß-Institut für
Angewandte Analysis und Stochastik

Premolab day Dec 2012

Vladimir Spokoiny

Outline

Data Y with DGP $Y \sim I\!P$.

PA: $I\!P \in (I\!P_\theta)$ or $I\!P = I\!P_{\theta^*}$ for $\theta^* \in \Theta$.

Maximum likelihood approach:

$$\widetilde{\theta} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta \in \Theta} L(\theta), \qquad L(\theta) = L(Y, \theta) = \log \frac{dI\!P_\theta}{d\mu}(Y)$$

If PA is probably wrong (PW), then $\widetilde{\theta}$ is still meaningful (quasi MLE), the target is defined by

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

Aim: inference on $\widetilde{\theta}$.

A linear model:

$$Y = \Psi^\top \theta + \varepsilon$$

with $Y, \varepsilon \in \mathbb{R}^n$, $\theta \in \mathbb{R}^p$, and Ψ - $p \times n$ design (feature) matrix.

$\varepsilon \sim \mathcal{N}(0, \Sigma^{-1})$ leads to a quasi MLE

$$\widetilde{\theta} = \operatorname*{argmax}_\theta L(\theta) = \operatorname*{argmin}_\theta (Y - \Psi^\top \theta)^\top \Sigma (Y - \Psi^\top \theta)$$

Gauss:

$$\widetilde{\theta} = D^{-2} \Psi \Sigma Y, \qquad D^2 \stackrel{\text{def}}{=} \Psi \Sigma \Psi^\top.$$

Consider PA $Y \sim \mathcal{N}(\Psi^\top \theta^*, \Sigma^{-1})$ yielding $L(\theta) = -(Y - \Psi^\top \theta)^\top \Sigma (Y - \Psi^\top \theta)/2 + R$

$$\widetilde{\theta} = \operatorname{argmax} L(\theta) = D^{-2} \Psi \Sigma Y, \qquad D^2 \stackrel{\text{def}}{=} \Psi \Sigma \Psi^\top .$$

Theorem (Gauss-Markov)

Assume a homogeneous noise $\mathbb{E}\varepsilon \equiv 0$, $\operatorname{Var}(\varepsilon) = \Sigma$. Then it holds for $\xi \stackrel{\text{def}}{=} D(\widetilde{\theta} - \theta^*)$

1. $\mathbb{E}\xi \equiv 0$ *and* $\operatorname{Var}(\xi) = I_p$;
2. $\widetilde{\theta}$ *is linearly efficient;*
3. $L(\widetilde{\theta}) - L(\theta) \equiv \|D(\widetilde{\theta} - \theta)\|^2/2$ *for any* θ;
4. $L(\widetilde{\theta}) - L(\theta^*) \equiv \|\xi\|^2/2$.

If PA is correct, i.e. $Y \sim \mathcal{N}(\theta^*, I_p)$, then

1. ξ *is standard normal;*
2. $2L(\widetilde{\theta}, \theta^*) \sim \chi^2_p$.

The first results 1-3 are entirely based on quadraticity of $L(\theta)$.

The results for the LR $L(\widetilde{\theta},\theta) = L(\widetilde{\theta}) - L(\theta)$

$$2L(\widetilde{\theta},\theta) = \|D(\widetilde{\theta}-\theta)\|^2, \qquad 2L(\widetilde{\theta},\theta^*) \sim \chi^2_p$$

can be used to build likelihood-based confidence ellipsoids for the parameter θ^*. Given $\mathfrak{z} > 0$, define

$$\mathcal{E}(\mathfrak{z}) = \{\theta : L(\widetilde{\theta},\theta) \le \mathfrak{z}\} = \{\theta : \|D(\widetilde{\theta}-\theta)\|^2 \le 2\mathfrak{z}\}. \tag{1}$$

Theorem

Assume $Y = \Psi^\top \theta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\Sigma)$ and consider the MLE $\widetilde{\theta}$. Define $\mathfrak{z}_\alpha$ by $P(\chi^2_p > 2\mathfrak{z}_\alpha) = \alpha$. Then $\mathcal{E}(\mathfrak{z}_\alpha)$ from (1) is an α-confidence set for θ^*.

Challenges

- Large p, in particular, $p \gg n$
- Poor identification with $D^2 = \Psi \Sigma \Psi^\top$ badly conditioned
- Model misspecification
- Semiparametric (functional) estimation

υ-setup:

$$Y = \Upsilon^\top \upsilon^* + \varepsilon = \Psi^\top \theta^* + \Phi^\top \eta^* + \varepsilon, \qquad \mathbb{E}\varepsilon = 0, \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Target: $\theta^* = P\upsilon^*$.

Profile qMLE 1: $\widetilde{\theta} = P\widetilde{\upsilon} = P(\Upsilon\Upsilon^\top)^{-1}\Upsilon Y = SY, \qquad S = P(\Upsilon\Upsilon^\top)^{-1}\Upsilon.$

Profile qMLE 2: $\widetilde{\theta} \stackrel{\mathrm{def}}{=} \operatorname*{argmax}_\theta \check{L}(\theta), \qquad \check{L}(\theta) \stackrel{\mathrm{def}}{=} \sup_{\upsilon:\, P\upsilon=\theta} L(\upsilon).$

Theorem

$$\mathbb{E}\widetilde{\theta} = \theta^*,$$
$$\mathrm{Var}(\widetilde{\theta}) = S\mathrm{Var}(\varepsilon)S^\top = \sigma^2 SS^\top = \sigma^2 P(\Upsilon\Upsilon^\top)^{-1}P^\top.$$

Model:

$$Y = \Psi^\top \theta^* + \Phi^\top \eta^* + \varepsilon \qquad \mathbb{E}\varepsilon = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Theorem

The profile MLE $\widetilde{\theta}$ reads as

$$\widetilde{\theta} = \left(\breve{\Psi}\breve{\Psi}^\top\right)^{-1}\breve{\Psi}Y,$$

$$\breve{\Psi} = \Psi - \Psi\Pi_\eta = \Psi - \Psi\Phi^\top\left(\Phi\Phi^\top\right)^{-1}\Phi.$$

Model:

$$Y = \Upsilon^\top \upsilon^* + \varepsilon = \Psi^\top \theta^* + \Phi^\top \eta^* + \varepsilon \qquad I\!E\varepsilon = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Theorem (Gauss-Markov)

1. $\widetilde{\theta} = SY$ *with* $S = P\left(\Upsilon\Upsilon^\top\right)^{-1}\Upsilon$ *fulfills*

$$\begin{aligned} I\!E\widetilde{\theta} &= \theta^* = P\upsilon^*, \\ I\!E\|\widetilde{\theta} - \theta^*\|^2 &= \mathrm{Var}(\widetilde{\theta}) = \sigma^2 P\left(\Upsilon\Upsilon^\top\right)^{-1} P^\top = \sigma^2\left(\breve{\Psi}\breve{\Psi}^\top\right)^{-1}, \\ \breve{\Psi} &= \Psi - \Psi\Pi_\eta \\ \Pi_\eta &= \Phi^\top\left(\Phi\Phi^\top\right)^{-1}\Phi. \end{aligned}$$

2. This risk is minimal in the class of all unbiased linear estimates of θ^* *.*

Define a CS for θ^* as the level set of $\check{L}(\theta) = \sup_{\upsilon: P\upsilon=\theta} L(\upsilon)$:

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \{\theta : \check{L}(\widetilde{\theta}) - \check{L}(\theta) \le \mathfrak{z}\}.$$

This definition can be rewritten as

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \Big\{\theta : \sup_{\upsilon} L(\upsilon) - \sup_{\upsilon: P\upsilon=\theta} L(\upsilon) \le \mathfrak{z}\Big\}.$$

In (θ, η)-setup the definition is even more transparent:

$$\mathcal{E}(\mathfrak{z}) \stackrel{\text{def}}{=} \Big\{\theta : \sup_{\theta,\eta} L(\theta,\eta) - \sup_{\eta} L(\theta,\eta) \le \mathfrak{z}\Big\}.$$

χ^2-Theorem

Model:

$$Y = \Psi^\top \theta^* + \Phi^\top \eta^* + \varepsilon \qquad \mathbb{E}\varepsilon = 0,\ \mathrm{Var}(\varepsilon) = \sigma^2 I_n.$$

Define

$$D^2 = \sigma^{-2} \breve{\Psi} \breve{\Psi}^\top, \qquad \breve{\Psi} = \Psi - \Psi \Pi_\eta .$$

Theorem

Let the matrix D be non-degenerated. It holds

$$\begin{aligned} 2\{\breve{L}(\widetilde{\theta}) - \breve{L}(\theta^*)\} &= \|D(\widetilde{\theta} - \theta^*)\|^2 = \|\xi\|^2, \\ \xi &= D(\widetilde{\theta} - \theta^*), \qquad \mathbb{E}\xi = 0,\ \mathrm{Var}(\xi) = I_p . \end{aligned}$$

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then ξ is standard normal in $\mathbb{R}^p$ and

$$2\{\breve{L}(\widetilde{\theta}) - \breve{L}(\theta^*)\} \sim \chi^2_p .$$

Let $\Sigma = \boldsymbol{I}_n$ and $\widetilde{\theta} = \left(\Psi\Psi^\top\right)^{-1}\Psi Y$.
Let R be a positive symmetric $p \times p$ matrix. Then the sum $\Psi\Psi^\top + R$ is positive symmetric as well and can be inverted whatever Ψ is.

This suggests to replace $\left(\Psi\Psi^\top\right)^{-1}$ by $\left(\Psi\Psi^\top + R\right)^{-1}$ leading to the regularized least squares estimate $\widetilde{\theta}_R$ of the parameter vector θ:

$$\widetilde{\theta}_R \stackrel{\text{def}}{=} \left(\Psi\Psi^\top + R\right)^{-1}\Psi Y,$$

Such a method is also called *ridge regression*.

Tikhonov regularization: $R = \alpha \boldsymbol{I}_p$ where $\alpha > 0$ and $\boldsymbol{I}_p$ stands for the unit matrix:

$$\widetilde{\theta}_\alpha \stackrel{\text{def}}{=} \left(\Psi\Psi^\top + \alpha \boldsymbol{I}_p\right)^{-1}\Psi Y.$$

MLE: with $L(\theta) = \sigma^{-2}\|Y - \Psi^\top \theta\|^2/2$

$$\widetilde{\theta} = \operatorname*{argmax}_\theta L(\theta) = \operatorname*{arginf}_\theta \|Y - \Psi^\top \theta\|^2$$

Roughness or complexity penalization: add a quadratic penalty $\|G\theta\|^2/2$ for a given symmetric matrix G:

$$\begin{aligned} L_G(\theta) &\stackrel{\text{def}}{=} L(\theta) - \|G\theta\|^2/2 \\ &= -(2\sigma^2)^{-1}\|Y - \Psi^\top \theta\|^2 - \|G\theta\|^2/2 - R. \end{aligned} \tag{2}$$

The penalized ML problem reads as

$$\widetilde{\theta}_G = \operatorname*{argmax}_\theta L_G(\theta) = \operatorname*{argmin}_\theta \left\{ (2\sigma^2)^{-1}\|Y - \Psi^\top \theta\|^2 + \|G\theta\|^2/2 \right\}.$$

Straightforward calculus yield

$$\widetilde{\theta}_G \stackrel{\text{def}}{=} \left(\Psi\Psi^\top + \sigma^2 G^2\right)^{-1}\Psi Y = S_G Y. \tag{3}$$

$\widetilde{\theta}_G$ estimates the value θ_G defined by

$$\begin{aligned}
\theta_G &= \operatorname*{argmax}_\theta I\!E L_G(\theta) \\
&= \operatorname*{arginf}_\theta I\!E\left\{\|Y - \Psi^\top\theta\|^2 + \sigma^2\|G\theta\|^2\right\} \\
&= \left(\Psi\Psi^\top + \sigma^2 G^2\right)^{-1}\Psi\Psi^\top\theta^*
\end{aligned}$$

and $\theta_G \neq \theta^*$ unless $G = 0$. In other words, the penalized MLE $\widetilde{\theta}_G$ is biased.

Theorem

Let $\widetilde{\theta}_G$ *be a penalized MLE from* (3)*. Under noise homogeneity* $\mathrm{Var}(\varepsilon) = \sigma^2 \boldsymbol{I}_n$*, it holds* $\mathbb{E}\widetilde{\theta}_G = \theta_G$ *and*

$$\mathrm{Var}(\widetilde{\theta}_G) = \sigma^2 S_G S_G^\top = \sigma^2 \left(\Psi\Psi^\top + \sigma^2 G^2\right)^{-1} \Psi\Psi^\top \left(\Psi\Psi^\top + \sigma^2 G^2\right)^{-1}.$$

The bias $\|\theta_G - \theta^*\|$ *monotonously increases in* G^2 *while the variance monotonously decreases with the penalization* G*.*
Finally

$$\mathbb{E}\|\widetilde{\theta}_G - \theta^*\|^2 = \|\theta_G - \theta^*\|^2 + \sigma^2 \mathrm{tr}\left(S_G S_G^\top\right).$$

Example (Tikhonov regularization): $\mathbb{E}\widetilde{\theta}_\alpha = \theta_\alpha$ for $\theta_\alpha = \left(\Psi\Psi^\top + \alpha \boldsymbol{I}_p\right)^{-1} \Psi\Psi^\top \theta^*$, the bias $\|\theta_\alpha - \theta^*\|$ grows with the regularization parameter α.

Theorem

Let $L_G(\theta)$ be the penalized log-likelihood from (2)*. Then*

$$\begin{aligned} 2L_G(\widetilde{\theta}_G, \theta_G) &= \left(\widetilde{\theta}_G - \theta_G\right)^\top \left(\sigma^{-2}\Psi\Psi^\top + G^2\right)\left(\widetilde{\theta}_G - \theta_G\right) \\ &= \sigma^{-2}\varepsilon^\top \Pi_G \varepsilon \end{aligned}$$

with $\Pi_G = \Psi^\top \left(\Psi\Psi^\top + \sigma^2 G^2\right)^{-1}\Psi$.

Define the confidence set $\mathcal{E}_G(\mathfrak{z})$ for θ_G as

$$\mathcal{E}_G(\mathfrak{z}) \stackrel{\text{def}}{=} \{\theta : L_G(\widetilde{\theta}_G, \theta) \le \mathfrak{z}\}.$$

The definition implies the following result for the coverage probability:

$$I\!P\big(\mathcal{E}_G(\mathfrak{z}) \not\ni \theta_G\big) \le I\!P\big(L_G(\widetilde{\theta}_G, \theta_G) > \mathfrak{z}\big)$$

Theorem

Let $L_G(\theta)$ be the penalized log-likelihood from (2) and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Then it holds with $\mathrm{p}_G = \mathrm{tr}(\Pi_G)$ and $v_G^2 = 2\,\mathrm{tr}(\Pi_G^2)$ that

$$I\!P\big(2L_G(\widetilde{\theta}_G, \theta_G) > \mathrm{p}_G + (2v_G \mathrm{x}^{1/2}) \vee (6\mathrm{x})\big) \le \exp(-\mathrm{x}).$$

Challenges in linear modeling

The key issue is an identifiability. Particularly challenging is any inference for

- inverse problems;
- poor design;
- dimension p too high;

More specific questions:

- Choice of a penalty class (given by G). Examples:
  - Projection (spectral cut-off)
  - Tikhonov
  - Roughness
  - diagonal (spectral weighting)
- Choice of a tuning parameters or aggregation
- Critical dimension (cardinality of the set of models)

More penalization

Penalized MLE:

$$L(\theta,\lambda) \stackrel{\text{def}}{=} L(\theta) + \lambda \mathfrak{t}(\theta) \qquad \widetilde{\theta}(\lambda) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} L(\theta,\lambda).$$

Complexity:

$$\mathfrak{t}(\theta) = \|\theta\|_0 = \#\{\theta_j \neq 0\}.$$

Sparse:

$$\mathfrak{t}(\theta) = \|\theta\|_1 = \sum_j |\theta_j|$$

Challenges:

- choice of penalty
- choice of λ
- constrained models under structural assumptions

Outline

Global modeling

- Global parametric MLE $\widetilde{\theta} = \operatorname{argmax} L(\theta)$, e.g. GLM, median and quantile regression, etc.
- Sieve parametric MLE
- Roughness penalty, regularization
- Complexity and sparse penalty

Issues:

- Choice of the basic model $L(\theta)$ and modeling bias;
- Choice of the sequence of models (sieves);
- Data-driven choice of a tuning parameter;
- Choice of the class of penalties $\mathrm{t}(\theta)$;

Local parametric

Even a complex model structure admits a simple description in a proper vicinity of each point.

Local approach: fix a vicinity and approximate the model function by a simple parametric model in this vicinity.

Issues:

- Choice of a family for local approximation (e.g. polynomials of a fixed degree)
- Choice of the locality parameter (bandwidth):
  - global vs pointwise,
  - isotropic vs anisotropic
- Uniformity and confidence bands

Structural smoothing

Semiparametric structural assumptions:

- single index $f(x) = g(x^\top \theta)$
- multiple index $f(x) = g(x^\top \theta_1, \ldots, x^\top \theta_m)$
- additive $f(x) = f_1(x_1) + \ldots + f_d(x_d)$
- projection pursuit $f(x) = f_1(x^\top \theta_1) + \ldots + f_d(x^\top \theta_d)$.

Local structures, manifold learning: $\theta = \theta(x)$.

Outline

PA: $I\!P \in \left(I\!P_\theta, \theta \in \Theta \subseteq I\!R^p\right)$.

Classical parametric viewpoint:

- PA is exactly fulfilled
- sample size n large

Main results:

- MLE $\widetilde{\theta}$ is asymptotically efficient.
- Wilks phenomenon:

$$2L(\widetilde{\theta}) - 2L(\theta^*) \xrightarrow{w} \chi^2_p.$$

“Modern” parametric modeling

Classical assumptions:

PA1 $I\!P \in (I\!P_\theta, \theta \in \Theta \subset I\!R^p)$

PA2 $n \to \infty$ or n/p large.

Requirements to the “modern” theory:

1. addressing “model misspecification” and “small samples”
2. consistent with the classical asymptotic results including the efficiency bounds.
3. extendable to the nonparametric situation.

Main tools:

- Local approximation
- Concentration of measures, probability bounds

General approach

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta} L(\theta), \quad \theta^* = \operatorname*{argmax}_{\theta} I\!E L(\theta).$$

- Concentration in a local neighborhood $\Theta_0(\mathtt{r})$ of θ^*

$$I\!P\big(\widetilde{\theta} \notin \Theta_0(\mathtt{r})\big) \le C_1 \exp\{-C_2 \mathtt{r}\}$$

- Local linear approximation:
  Approximate $L(\theta)$ within $\Theta_0(\mathtt{r})$ by a quadratic (in θ) expansion:

$$I\!E L(\theta) - I\!E L(\theta^*) \approx -\frac{1}{2}\|D_0(\theta - \theta^*)\|^2, \qquad D_0^2 \stackrel{\text{def}}{=} \nabla^2 I\!E L(\theta^*),$$

$$\zeta(\theta) - \zeta(\theta^*) \approx (\theta - \theta^*)^\top \nabla\zeta(\theta^*), \qquad \zeta(\theta) \stackrel{\text{def}}{=} L(\theta) - I\!E L(\theta).$$

  Yields

$$L(\theta, \theta^*) \approx (\theta - \theta^*)^\top \nabla\zeta(\theta^*) - \frac{1}{2}\|D_0(\theta - \theta^*)\|^2$$

  Problem: error of approximation is too big.

A large deviation bound. Classic vs modern

Classical conditions:

- Θ is compact;
- identifiability:

$$I\!E L(\theta^*) - I\!E L(\theta) > 0.$$

- stochastic continuity of $L(\theta)$ and continuity of $I\!E L(\theta)$;
- large samples.

"Modern" approach:

- Local smoothness of $L(\theta)$: $\mathrm{Var}\{\nabla L(\theta)\} \le V^2$
- Qualified identifiability:

$$I\!E L(\theta^*) - I\!E L(\theta) \ge \begin{cases} \mathfrak{b}\|V(\theta-\theta^*)\|^2 & \text{or,} \\ C_1 p \log\bigl(\|V(\theta-\theta^*)\|^2/p\bigr) & . \end{cases}$$

LD bound. Conditions

Let $\zeta(\theta) = L(\theta) - I\!\!E L(\theta)$.

(ED) $\exists$ *a positive symmetric matrix* V^2, *and constants* $\mathrm{g} > 0$, $\nu_0 \geq 1$ *s.t. for all* $\theta \in \Theta$

$$\sup_{\gamma \in \mathcal{S}^p} \log I\!\!E \exp\left\{ \lambda \frac{\gamma^\top \nabla \zeta(\theta)}{\|V\gamma\|} \right\} \leq \nu_0^2 \lambda^2/2, \quad |\lambda| \leq \mathrm{g}.$$

$(\mathcal{L})$ *There is a constant* $C_1 \geq 1$ *s.t.*

$$\mathfrak{M}^*(\theta, \theta^*) \geq C_1 p^* \log\bigl(\|V(\theta - \theta^*)\|^2 / p^*\bigr),$$

where

$$\mathfrak{M}^*(\theta, \theta^*) \stackrel{\mathrm{def}}{=} \sup_{\mu \in \mathbb{M}} \mathfrak{M}(\mu, \theta, \theta^*).$$

Theorem

Suppose Then it holds for $\Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\theta : \|V_0(\theta - \theta^*)\| \le \mathtt{r}\}$

$$I\!P\big(\widetilde{\theta} \notin \Theta_0(\mathtt{r})\big) \le C_1 \exp\big\{-C_2\mathtt{r}\big\}.$$

- Ensures concentration of $\widetilde{\theta}$ in a vicinity of θ^* of radius $\asymp \sqrt{p/n}$.
- The bound non-asymptotic, applies for small samples and nearly sharp. All constants implicit.
- The true measure $I\!P$ does not belong to $(I\!P_\theta)$.

Tools:

- Generic chaining in a local ball; Talagrand (2005), Bednorz (2006);
- Slicing arguments

Local conditions

(ED_0) $\exists$ *a positive symmetric matrix* V_0^2, *and constants* $\mathtt{g} > 0$, $\nu_0 \geq 1$ *s.t.*

$$\sup_{\gamma \in \mathcal{S}^p} \log I\!\!E \exp\left\{ \lambda \frac{\gamma^\top \nabla \zeta(\theta^*)}{\|V_0 \gamma\|} \right\} \leq \nu_0^2 \lambda^2/2, \quad |\lambda| \leq \mathtt{g}.$$

(ED_1) *For some* R *and each* $\mathtt{r} \leq R$, *there exist a constant* $\omega(\mathtt{r}) \leq 1/2$ *such that it holds for all* $\theta \in \Theta_0(\mathtt{r}) = \Theta_0(\mathtt{r}) \stackrel{\text{def}}{=} \{\theta : \|V_0(\theta - \theta^*)\| \leq \mathtt{r}\}$:

$$\sup_{\gamma \in \mathcal{S}^p} \log I\!\!E \exp\left\{ \lambda \frac{\gamma^\top \{\nabla \zeta(\theta) - \nabla \zeta(\theta^*)\}}{\omega(\mathtt{r}) \|V_0 \gamma\|} \right\} \leq \nu_0^2 \lambda^2/2, \qquad |\lambda| \leq \mathtt{g}.$$

$(\mathcal{L}_0)$ *There are a matrix* D_0 *and for each* $\mathtt{r} \leq R$ *and a constant* $\delta(\mathtt{r}) \leq 1/2$, *such that it holds on the set* $\Theta_0(\mathtt{r}) = \{\theta : \|V_0(\theta - \theta^*)\| \leq \mathtt{r}\}$

$$\left| \frac{-2 I\!\!E L(\theta, \theta^*)}{\|D_0(\theta - \theta^*)\|^2} - 1 \right| \leq \delta(\mathtt{r}).$$

Local bracketing

Yields

$$L(\theta,\theta^*) \approx (\theta-\theta^*)^\top \nabla\zeta(\theta^*) - \frac{1}{2}\|D_0(\theta-\theta^*)\|^2$$

Problem: error of approximation is too big.

Approach: majorate from above and from below by log-likelihoods of two different linear models:

With $\varepsilon = (\delta,\rho)$, define

$$\mathbb{L}_\varepsilon(\theta,\theta^*) \stackrel{\text{def}}{=} (\theta-\theta^*)^\top \nabla L(\theta^*) - \|D_\varepsilon(\theta-\theta^*)\|^2/2$$

$$\mathbb{L}_{\underline{\varepsilon}}(\theta,\theta^*) \stackrel{\text{def}}{=} (\theta-\theta^*)^\top \nabla L(\theta^*) - \|D_{\underline{\varepsilon}}(\theta-\theta^*)\|^2/2$$

where with $V_0^2 = \mathrm{Var}\big(\nabla\zeta(\theta^*)\big)$

$$D_\varepsilon^2 = D_0^2(1-\delta) - \rho V_0^2,$$

$$D_{\underline{\varepsilon}}^2 = D_0^2(1+\delta) + \rho V_0^2,$$

The construction depends upon the unknown quantities θ^*, D_0, V_0.

Majorization bound

$$\mathbb{L}_{\varepsilon}(\theta,\theta^*) = (\theta-\theta^*)^{\top}\nabla L(\theta^*) - \|D_{\varepsilon}(\theta-\theta^*)\|^2/2$$

$$\mathbb{L}_{\underline{\varepsilon}}(\theta,\theta^*) = (\theta-\theta^*)^{\top}\nabla L(\theta^*) - \|D_{\underline{\varepsilon}}(\theta-\theta^*)\|^2/2,$$

$$D_{\varepsilon} = D_0^2(1-\delta) - \rho V_0^2, \qquad \varepsilon = (\delta,\rho), \qquad \underline{\varepsilon} = -\varepsilon = (-\delta,-\rho).$$

Theorem

Assume (ED_1) *and* $(\mathcal{L}_0)$*. Let for some* $\mathtt{r}$*, the values* $\rho \geq \sqrt{3\nu_0}\,\omega(\mathtt{r})$ *and* $\delta \geq \delta(\mathtt{r})$ *ensure* $D_{\varepsilon} \geq 0$*. Then*

$$\mathbb{L}_{\underline{\varepsilon}}(\theta,\theta^*) - \diamondsuit_{\underline{\varepsilon}}(\mathtt{r}) \leq L(\theta,\theta^*) \leq \mathbb{L}_{\varepsilon}(\theta,\theta^*) + \diamondsuit_{\varepsilon}(\mathtt{r}), \qquad \theta \in \Theta_0(\mathtt{r}),$$

where the random variables $\diamondsuit_{\varepsilon}(\mathtt{r}), \diamondsuit_{\underline{\varepsilon}}(\mathtt{r})$ *fulfill*

$$\log I\!E \exp\bigl\{\diamondsuit_{\varepsilon}(\mathtt{r})/\omega(\mathtt{r}) - \mathfrak{c}_1 p^*\bigr\} \leq 0.$$

Usually $\omega(\mathtt{r}) \asymp n^{-1/2}\mathtt{r}$ and $\delta(\mathtt{r}) \asymp n^{-1/2}\mathtt{r}$.

Comments

Tools:

- Generic chaining + slicing
- Sharp deviation bounds for quadratic forms of Gaussian and non-Gaussian r.v.

Issues:

- The bound is non-asymptotic, applies even for small samples
- The set $\Theta_0(\mathtt{r})$ should not be root-n small. It corresponds to a usual local vicinity.
- No normalization/scaling necessary, the bounds are self-normalizing.
- The difference $\mathbb{L}_\varepsilon(\theta,\theta^*) - \mathbb{L}_{\underline{\varepsilon}}(\theta,\theta^*)$ can be large. Matters "upper value" - "lower value":

$$\sup_\theta \mathbb{L}_\varepsilon(\theta,\theta^*) - \sup_\theta \mathbb{L}_{\underline{\varepsilon}}(\theta,\theta^*).$$

Local coverage probability

Consider the confidence set

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

Coverage probability:

$$I\!P\{\mathcal{E}(\mathfrak{z}) \not\ni \theta^*\} = I\!P\{L(\widetilde{\theta}, \theta^*) > \mathfrak{z}\}.$$

Theorem

For any $\mathfrak{z} > 0$, it holds with $\xi_\varepsilon = D_\varepsilon^{-1} \nabla L(\theta^*)$

$$I\!P\{\mathcal{E}(\mathfrak{z}) \not\ni \theta^*, \widetilde{\theta} \in \Theta_0(\mathtt{r})\} \leq I\!P\{\|\xi_\varepsilon\|^2 \geq 2\mathfrak{z} - 2\diamondsuit_\varepsilon(\mathtt{r})\}.$$

Define

$$\Delta_\varepsilon(\mathtt{r}) \stackrel{\text{def}}{=} \diamondsuit_\varepsilon(\mathtt{r}) + \diamondsuit_{\underline{\varepsilon}}(\mathtt{r}) + \bigl(\|\xi_\varepsilon\|^2 - \|\xi_{\underline{\varepsilon}}\|^2\bigr)/2.$$

Theorem

For any $z > 0$, it holds

$$I\!P\bigl\{\|D_\varepsilon(\widetilde{\theta} - \theta^*)\| > z, \widetilde{\theta} \in \Theta_0(\mathtt{r})\bigr\} \le I\!P\bigl\{\|\xi_\varepsilon\| > z - \sqrt{2\Delta_\varepsilon(\mathtt{r})}\bigr\}.$$

Theorem (Wilks phenomenon)

The following approximation holds on the random set $C_\varepsilon(\mathtt{r}) = \{\widetilde{\theta} \in \Theta_0(\mathtt{r}), \|\xi_{\underline{\varepsilon}}\| \le \mathtt{r}\}$ *:*

$$\|\xi_{\underline{\varepsilon}}\|^2/2 - \diamondsuit_{\underline{\varepsilon}}(\mathtt{r}) \le L(\widetilde{\theta}, \theta^*) \le \|\xi_\varepsilon\|^2/2 + \diamondsuit_\varepsilon(\mathtt{r}).$$

Theorem (Fisher expansion)

On the same random set $C_\varepsilon(\mathtt{r})$

$$\left\|D_\varepsilon\left(\widetilde{\theta} - \theta^*\right) - \xi_\varepsilon\right\|^2 \le 2\Delta_\varepsilon(\mathtt{r}).$$

Critical dimension

- The results are useless if the error Δ_ε is too big.
- The value Δ_ε depends on the dimension p.
- Critical dimension p_n ?

Theorem

In the regular i.i.d. case the Wilks result holds if

$$p_n/n \text{ is small.}$$

Outline

Regularization and Penalization

Modern applications:

- large parameter space,
- lack of identifiability.

Approach: penalization:

$$\widetilde{\theta} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} \{L(\theta) - \mathfrak{t}(\theta)\}$$

- smooth (roughness) penalty $\mathfrak{t}(\theta) = \lambda \|G\theta\|^2/2$
- complexity penalty $\mathfrak{t}(\theta) = \lambda \|\theta\|_0$
- sparcity penalty $\mathfrak{t}(\theta) = \lambda \|\theta\|_1$

New target:

$$\theta^*_G \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} \{I\!\!E L(\theta) - \mathfrak{t}(\theta)\}.$$

The general theory extends easily to the smooth and complexity penalization.

Challenge: a general theory for the sparse case.

Choice of a regularization parameter

$$\widetilde{\theta} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta} \{L(\theta) - \mathfrak{t}(\theta)\} = \operatorname*{argmax}_{\theta} \{L(\theta) - \lambda \|G\theta\|^2/2\}$$

Choice of λ crucial (central problem in nonparametric statistics).

Approach: include λ in the list of parameters.

Naive definition

$$(\widetilde{\theta}, \widetilde{\lambda}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta,\lambda} \{L(\theta) - \lambda \|G\theta\|^2/2\}$$

Does not work (yields overfitting). Better

$$(\widetilde{\theta}, \widetilde{\lambda}) \stackrel{\text{def}}{=} \operatorname*{argmax}_{\theta,\lambda} \{L(\theta) - \lambda \|G\theta\|^2/2 - \mathfrak{t}(\lambda)\}$$

New challenge: choice of $\mathfrak{t}(\lambda)$.

Examples: AIC, BIC, SCAD, ...

Bayes approach

$$Y \mid \theta \sim p(y|\theta),$$

$$\vartheta \sim \pi(\cdot).$$

ϑ random from the prior π.

Finally the *posteriori (conditional) density* $p(\vartheta|y)$ of ϑ given y:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(y|\theta)\pi(\theta)}{\int_\Theta p(y|\theta)\pi(\theta)\lambda(d\theta)}.$$

or

$$\vartheta \mid Y \propto \exp\{L(\theta)\}\,\Pi(d\theta).$$

Similarly to out general qMLE approach

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

Aim: concentration of the posterior, near Gaussianity.

Bernstein - von Mises Theorem

- Non-informative prior: $\pi(d\theta) = d\theta$.

Posterior $\vartheta \,|\, Y$ fulfills:

$$I\!E\big(\vartheta \,\big|\, Y\big) \approx \widetilde{\theta} \qquad \mathrm{Var}\big(\vartheta \,\big|\, Y\big) \approx D_0^{-2}, \qquad D_0\big(\vartheta \,\big|\, Y\big) \approx \xi \sim \mathcal{N}(0, \boldsymbol{I}_p)$$

- Gaussian priors: $d\pi_G(\theta) \propto \exp\big(-\|G\theta\|^2/2\big)$.

Posterior $\vartheta_G \,|\, Y$ fulfills:

$$I\!E\big(\vartheta_G \,\big|\, Y\big) \approx \widetilde{\theta}_G \qquad \mathrm{Var}\big(\vartheta \,\big|\, Y\big) \approx D_G^{-2}, \qquad D_G\big(\vartheta_G \,\big|\, Y\big) \approx \xi \sim \mathcal{N}(0, \boldsymbol{I}_p)$$

- Critical dimension: the BvM result applies if "p/n is small".

More BvM: New challenges

Challenge 1: Sparse priors: (with heavy tails). Only few special cases studied.

Challenge 2: Bayesian model selection and Bernstein - von Mises Theorem for hyperpriors (priors on prior parameters).

Data Y with DGP $Y \sim I\!P$.

SPA: $I\!P \in (I\!P_{\theta,\eta})$, probably misspecified.

θ, target, η, nuisance.

Goal: inference on θ.

Examples in mind:

- an inverse problem with error in operator;
observed Y and $\widehat{A}$ following $Y = A\theta + \varepsilon$ with θ and A unknown;

- IV problem; see e.g. Hall and Horowitz (2005), Horowitz (2011)

- dimension reduction in single- and multi-index models; e.g. Xia et al (2000), Hristache et al (2001a,b).

Profile MLE

SPA : $Y \sim I\!P \in \left(I\!P_{\theta,\eta}, \theta \in \Theta, \eta \in H\right)$

Log-likelihood: $L(\theta,\eta) = \dfrac{dI\!P_{\theta,\eta}}{d\mu_0}(Y)$

Profile MLE: $\widetilde{\theta} = \operatorname*{argmax}_{\theta} \max_{\eta} L(\theta,\eta) = \operatorname*{argmax}_{\theta} \check{L}(\theta), \qquad \check{L}(\theta) = \max_{\eta} L(\theta,\eta).$

Murphy, van der Vaart (2000), Kosorok (2005, 2008): Let $I\!P = I\!P_{\theta^*,\eta^*}$. Then $\widetilde{\theta}$ is

- root-n consistent and normal
- semiparametrically efficient
- $2\check{L}(\widetilde{\theta}) - 2\check{L}(\theta^*) \xrightarrow{w} \chi^2_p$, where $p = \dim(\Theta)$.

Limitations:

- hard optimization problem, often unfeasible
- SPA is crucial but questionable
- large sample asymptotics

Iterative methods

- Alternating or EM-algorithms: iterate two steps until convergence;
- backfitting / projection pursuit;
- MCMC, Gibbs, . . .

Problems:

- convergence, asymptotic properties, efficiency only addressed in few cases;
- no general theory available;

Challenges: a general study for an alternating method focusing on:

- finite samples;
- model misspecification;
- unified approach for direct and inverse problems;
- range of applicability, critical dimension of the nuisance;

Outline

Dimension-reduction

Observed $Y \in \mathbb{R}^1, X \in \mathbb{R}^d$, d large.
Aim: regression $Y = f(X) + \varepsilon$. Unfeasible for $d \geq 3$.
EDR assumptions:

- Single-index:

$$f(X) = g(X^\top \theta).$$

- Multi-index:

$$f(X) = g(TX), \qquad T : \mathbb{R}^d \to \mathbb{R}^m$$

- Additive

$$f(X) = g_1(X_1) + \ldots + g_d(X_d)$$

- Projection-pursuit

$$f(X) = g_1(X^\top \theta_1) + \ldots + g_d(X^\top \theta_d)$$

Analysis of a high-dimensional distribution

$X_1, \ldots, X_n \in \mathbb{R}^d$, i.i.d. from $p(\cdot)$

Aim: features of $p(\cdot)$ like multimodality or heavy tails.

Non-Gaussian components:

$$p(x) = g(TX)\phi_\Gamma(x)$$

T , a non-Gaussian subspace.

Applications in clustering (unsupervised learning) for multimodality and in risk management in heavy tails

Error in operator problems

Linear inverse problem:

$$Y = A\theta + \varepsilon$$

A unknown, instead a pilot $\widehat{A}$ is available.

Aim: estimation of θ, the operator A is the nuisance parameter.

Applications in Instrumental Variable regression, functional data analysis, imaging, and many others

Conditional moment restrictions

CMR modeling:

$$I\!E\left[g(Z,\theta)\,|\,X\right]=0 \quad \text{a.s.} \quad \Leftrightarrow \quad \theta=\theta^*.$$

Approach:

$$\widetilde{\theta}=\widetilde{\theta}_h \stackrel{\text{def}}{=} \operatorname*{argmin}_{\theta\in\Theta} M(\theta)$$

with

$$M(Z,\theta) \stackrel{\text{def}}{=} \sum_{i\neq j} g_i(\theta)g_j(\theta)w_{ij}$$

where w_{ij} are localizing weights with the bandwidth h.

Challenge: Theory for quadratic forms.

More topics

- Adaptive clustering
- Adaptive classification
- Random design and time series
- Structure adaptive manifold smoothing (locally adaptive dimension reduction).
- Nonparametric change point detection
- Hidden Markov chain modeling
- Resampling methods