Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz-Gemeinschaft

Fisher and Wilks expansions with applications to statistical inference

Vladimir Spokoiny,
WIAS, HU Berlin

Outline

1 **Bernstein – von Mises Theorem**
- BvM Theorem
- Local Gaussian approximation of the posterior
- Tail posterior probability and contraction
- Credible sets

2 **Semiparametric estimation**
- Motivation
- Linear models
- General semiparametric setup

3 **Penalized MLE and effective dimension**
- Curse of dimension
- Effective dimension
- Fisher and Wilks expansions
- Concentration and large deviations
- A bound for the norm of a vector stochastic process

4 **Confidence estimation using bootstrap**
- Likelihood-based confidence set
- Multipier bootstrap
- Conditions

Let $\boldsymbol{\vartheta}$, a random element Θ,

$\pi(\boldsymbol{\theta})$ a prior density.

The posterior distribution of $\boldsymbol{\vartheta}$ is given by

$$\mathbb{P}(A \,\big|\, \boldsymbol{Y}) = \frac{\int_A \exp\bigl\{L(\boldsymbol{\theta})\bigr\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int_\Theta \exp\bigl\{L(\boldsymbol{\theta})\bigr\}\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

Introduce the posterior moments

$$\overline{\boldsymbol{\vartheta}} \stackrel{\text{def}}{=} \mathbb{E}\bigl(\boldsymbol{\vartheta} \,\big|\, \boldsymbol{Y}\bigr),$$

$$\mathfrak{S}^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\boldsymbol{\vartheta} \,\big|\, \boldsymbol{Y}) \stackrel{\text{def}}{=} \mathbb{E}\bigl\{(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})(\boldsymbol{\vartheta} - \overline{\boldsymbol{\vartheta}})^\top \,\big|\, \boldsymbol{Y}\bigr\}.$$

Some references

There is a number of papers in this direction recently appeared:

- [Ghosal et al., 2000, Ghosal and van der Vaart, 2007] for a general theory in the i.i.d. case;
- [Ghosal, 1999], [Ghosal, 2000] for high dimensional linear models;
- [Boucheron and Gassiat, 2009], [Kim, 2006] for some special non-Gaussian models;
- [Shen, 2002], [Bickel and Kleijn, 2012], [Rivoirard and Rousseau, 2012], [Castillo, 2012], [Castillo and Rousseau, 2013] for a semiparametric version of the BvM result for different models;
- [Kleijn and van der Vaart, 2006], [Bunke and Milhaud, 1998], for the misspecified parametric case,
- [Castillo and Rousseau, 2013],
- [Kleijn and van der Vaart, 2012] for a general framework for the BvM result in terms of a stochastic LAN condition

Extensions to nonparametric models with infinite or growing parameter dimension p exist for some special situations:

- [Freedman, 1999] and [Ghosal, 1999, Ghosal, 2000] for linear models
- [Bontemps, 2011] for Gaussian regression,
- [Castillo and Nickl, 2013] for the white noise case;

Below $\pi(\boldsymbol{\theta}) \equiv 1$, an improper non-informative prior.

Yields for any $A \subset \Theta$

$$I\!P^{\circ}(A) = I\!P(A \,|\, \boldsymbol{Y}) = \frac{\int_A \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}}{\int_{\Theta} \exp\{L(\boldsymbol{\theta})\} d\boldsymbol{\theta}} \, .$$

Quasi-likelihood $\Longrightarrow$ quasi-posterior.

A general case with a continuous prior density $\pi(\boldsymbol{\theta})$:

$$\boldsymbol{\vartheta} \,|\, \boldsymbol{Y} \propto \exp\{L(\boldsymbol{\theta})\}\pi(\boldsymbol{\theta}) = \exp\{L_{\pi}(\boldsymbol{\theta})\}$$

with

$$L_{\pi}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) + \log \pi(\boldsymbol{\theta}).$$

So, the case of a general smooth prior can be reduced to the case of a non-informative prior by changing the log-likelihood function.

Theorem

Suppose the conditions of Theorem 19. Let also $\mathtt{b}(\mathtt{r})$ *from* $(\mathcal{L})$ *satisfies*

$$\mathtt{r}^2 \mathtt{b}^2(\mathtt{r}) \geq \mathtt{x} + 2p + 4z^2(B, \mathtt{x}) + 8\mathtt{r}\, \mathtt{b}(\mathtt{r}) \varrho(\mathtt{r}, \mathtt{x}), \qquad \mathtt{r} \geq \mathtt{r}_0, \tag{1}$$

with $\varrho(\mathtt{r}, \mathtt{x})$ *from* (14). *Then it holds on a random set* $\Omega(\mathtt{x})$ *of probability at least* $1 - 5\mathrm{e}^{-\mathtt{x}}$

$$I\!\!P\bigl(\boldsymbol{\vartheta} \notin \Theta_0(\mathtt{r}_0) \,\big|\, \boldsymbol{Y}\bigr) \leq \mathrm{e}^{-\mathtt{x}}.$$

The bound (1) is very similar to the bound for the MLE concentration. It can be spelled out as the condition that

- $\mathtt{r}_0^2 \geq 2p + \mathtt{x} + 4\mathfrak{z}^2(B, \mathtt{x})$,
- $\mathtt{b}(\mathtt{r}_0) \approx 1$, and
- $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows with $\mathtt{r}$.

Define

$$\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D_0^{-2} \nabla L(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^* + D_0^{-1} \boldsymbol{\xi}.$$

The Fisher result implies

$$\| D_0 (\widetilde{\boldsymbol{\theta}} - \breve{\boldsymbol{\theta}}) \| \leq \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

Theorem

On $\Omega(\mathtt{x})$

$$\| D_0 (\overline{\boldsymbol{\vartheta}} - \breve{\boldsymbol{\theta}}) \|^2 \leq 4 \Delta(\mathtt{r}_0, \mathtt{x}) + 4 \mathrm{e}^{-\mathtt{x}},$$

$$\left\| I_p - D_0 \mathfrak{S}^2 D_0 \right\|_\infty \leq 4 \Delta(\mathtt{r}_0, \mathtt{x}) + 4 \mathrm{e}^{-\mathtt{x}}.$$

$$\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D_0^{-2} \nabla L(\boldsymbol{\theta}^*) = \boldsymbol{\theta}^* + D_0^{-1} \boldsymbol{\xi}.$$

Theorem

For any $\boldsymbol{\lambda} \in I\!R^p$ *with* $\|\boldsymbol{\lambda}\|^2 \le p$

$$\left| \log I\!E \left[\exp\{ \boldsymbol{\lambda}^\top D_0 (\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \} \,\middle|\, \boldsymbol{Y} \right] - \|\boldsymbol{\lambda}\|^2 / 2 \right| \le 2\Delta(\mathtt{r}_0, \mathtt{x}) + 3\mathrm{e}^{-\mathtt{x}},$$

and for any measurable set $A \subset I\!R^p$

$$I\!P\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \in A \,\big|\, \boldsymbol{Y} \bigr) \ge \exp\bigl\{ -2\Delta(\mathtt{r}_0, \mathtt{x}) - 3\mathrm{e}^{-\mathtt{x}} \bigr\} I\!P\bigl(\boldsymbol{\gamma} \in A \bigr) - \mathrm{e}^{-\mathtt{x}},$$

$$I\!P\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \in A \,\big|\, \boldsymbol{Y} \bigr) \le \exp\bigl\{ 2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}} \bigr\} I\!P\bigl(\boldsymbol{\gamma} \in A \bigr) + \mathrm{e}^{-\mathtt{x}}.$$

- All statements of Theorem 1 require " $\varDelta(\mathtt{r}_0, \mathtt{x})$ is small".

- The BvM result is stated under essentially the same list of conditions as the frequentist results of Fisher and Wilks Theorems.

- The normal approximation of the posterior is entirely based on the smoothness properties of the likelihood function

- No any asymptotic arguments like weak convergence or convergence in probability, or the Central Limit Theorem.

- The results continue to hold if $\breve{\boldsymbol{\theta}}$ is replaced by any efficient estimate $\widehat{\boldsymbol{\theta}}$, e.g. by the MLE $\widetilde{\boldsymbol{\theta}}$, satisfying $\|D_0(\widehat{\boldsymbol{\theta}} - \breve{\boldsymbol{\theta}})\| \le \mathtt{r}_0$ with a dominating probability.

Steps: Local Gaussian approximation of the posterior

Remind $D_0^2 = -\nabla^2 I\!\!E L(\boldsymbol{\theta}^*)$, $\boldsymbol{\xi} = D_0^{-1} \nabla L(\boldsymbol{\theta}^*)$, and

$$\breve{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D_0^{-1}\boldsymbol{\xi} = \boldsymbol{\theta}^* + D_0^{-2}\nabla L(\boldsymbol{\theta}^*)$$

Local approximation: on $\Omega(\mathtt{x})$, for $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - \frac{1}{2}\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2$

$$\bigl| L(\boldsymbol{\theta},\boldsymbol{\theta}^*) - \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \bigr| \le \Delta(\mathtt{r}_0,\mathtt{x}), \quad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0). \tag{2}$$

For any nonnegative function f , it holds

$$\int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{ L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \bigr\} f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta}$$

$$\le \mathrm{e}^{\Delta(\mathtt{r}_0,\mathtt{x})} \int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{ \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \bigr\} f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta}.$$

$$\int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{ L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \bigr\} f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta}$$

$$\ge \mathrm{e}^{-\Delta(\mathtt{r}_0,\mathtt{x})} \int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{ \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \bigr\} f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta}.$$

The main benefit: $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ is quadratic in $\boldsymbol{\theta}$ and thus

$$\exp \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = \exp\bigl\{ \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - \| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \|^2 / 2 \bigr\}$$

is proportional to the density of a Gaussian distribution.

More precisely, define

$$m(\boldsymbol{\xi}) \stackrel{\text{def}}{=} -\|\boldsymbol{\xi}\|^2/2 + \log(\det D_0) - p \log(\sqrt{2\pi}).$$

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = -\| D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}}) \|^2/2 + \log(\det D_0) - p \log(\sqrt{2\pi}) \qquad (3)$$

is (conditionally on $\boldsymbol{Y}$) the log-density of the normal law $\mathcal{N}(\breve{\boldsymbol{\theta}}, D_0^{-2})$ with the mean $\breve{\boldsymbol{\theta}} = D_0^{-1} \boldsymbol{\xi} + \boldsymbol{\theta}^*$ and the covariance matrix D_0^{-2} .

Theorem

For any nonnegative function $f(\cdot)$ *on* $I\!R^p$, *it holds on* $\Omega(\mathtt{r}_0, \mathtt{x})$

$$I\!E^{\circ}\bigl[f\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}})\bigr)\, 1\!\!\mathrm{I}_{\mathtt{r}_0}\bigr] \le \exp\bigl\{\Delta^{+}(\mathtt{r}_0, \mathtt{x})\bigr\}\, I\!E f(\boldsymbol{\gamma}), \tag{4}$$

where

$$\Delta^{+}(\mathtt{r}_0, \mathtt{x}) = 2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0),$$

$$\nu(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} -\log I\!P^{\circ}\bigl(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \le \mathtt{r}_0\bigr).$$

If $\mathtt{r}_0^2 \ge z^2(B, \mathtt{x}) + p + 2\mathtt{x}$, *then on* $\Omega(B, \mathtt{x})$, *it holds*

$$\nu(\mathtt{r}_0) \le 2\mathrm{e}^{-\mathtt{x}}$$

$$\Delta^{+}(\mathtt{r}_0, \mathtt{x}) \le 2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}}.$$

We use that $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) = \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - \|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\|^2/2$ is proportional to the density of a Gaussian distribution. More precisely, define

$$m(\boldsymbol{\xi}) \stackrel{\text{def}}{=} -\|\boldsymbol{\xi}\|^2/2 + \log(\det D_0) - p\log(\sqrt{2\pi}).$$

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*) \;=\; -\|D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\|^2/2 + \log(\det D_0) - p\log(\sqrt{2\pi}) \tag{5}$$

is (conditionally on $\boldsymbol{Y}$) the log-density of the normal law with the mean $\breve{\boldsymbol{\theta}} = D_0^{-1}\boldsymbol{\xi} + \boldsymbol{\theta}^*$ and the covariance matrix D_0^{-2} . Change of variables $\boldsymbol{u} = D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})$ implies by (5) for any nonnegative function f that

$$\begin{aligned}
&\int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*) + m(\boldsymbol{\xi})\bigr\}\, f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta} \\
&\quad\le \mathrm{e}^{\Delta(\mathtt{r}_0,\mathtt{x})} \exp\bigl\{m(\boldsymbol{\xi})\bigr\} \int \exp\bigl\{\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\}\, f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta} \\
&\quad= \mathrm{e}^{\Delta(\mathtt{r}_0,\mathtt{x})} \int \phi(\boldsymbol{u})\, f(\boldsymbol{u})\, d\boldsymbol{u} = \mathrm{e}^{\Delta(\mathtt{r}_0,\mathtt{x})}\, I\!E f(\boldsymbol{\gamma}).
\end{aligned} \tag{6}$$

Similarly, for any nonnegative function f, it follows by change of variables $\boldsymbol{u} = D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})$ and $D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \boldsymbol{u} + \boldsymbol{\xi}$ that

$$\exp\{m(\boldsymbol{\xi})\} \int \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} f\bigl(D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\bigr) \, \mathbb{1}\{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}_0\} d\boldsymbol{\theta}$$

$$\ge \exp\{-\varDelta(\mathtt{r}_0, \mathtt{x})\} \int \phi(\boldsymbol{u}) f(\boldsymbol{u}) \, \mathbb{1}\{\|\boldsymbol{u} + \boldsymbol{\xi}\| \le \mathtt{r}_0\} d\boldsymbol{u}. \tag{7}$$

A special case of (7) with $f(\boldsymbol{u}) \equiv 1$ implies by definition of $\nu(\mathtt{r}_0)$:

$$\exp\{m(\boldsymbol{\xi})\} \int_{\varTheta_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} \, d\boldsymbol{\theta} \ge \exp\bigl\{-\varDelta(\mathtt{r}_0, \mathtt{x}) - \nu(\mathtt{r}_0)\bigr\}. \tag{8}$$

Now we are prepared to finalize the proof. (6) and (8) imply on $\Omega(\mathtt{r}_0, \mathtt{x})$

$$\frac{\int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} f\bigl(D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\bigr) \, d\boldsymbol{\theta}}{\int \exp\bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} \, d\boldsymbol{\theta}} \le \exp\bigl\{2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0)\bigr\} \, I\!\!E f(\boldsymbol{\gamma})$$

and (4) follows. As $\|\boldsymbol{\xi}\| \le z(B, \mathtt{x})$ on $\Omega(B, \mathtt{x})$ and $\mathtt{r}_0 \ge z(B, \mathtt{x}) + z(p, \mathtt{x})$,

$$\nu(\mathtt{r}_0) = -\log I\!\!P^{\circ}\bigl(\|\boldsymbol{\gamma} + \boldsymbol{\xi}\| \le \mathtt{r}_0\bigr) \le -\log I\!\!P\bigl(\|\boldsymbol{\gamma}\| \le z(p, \mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}},$$

Lemma

For each $\mathtt{x}$ *and for* $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$

$$I\!\!P\bigl(\|\boldsymbol{\gamma}\| \ge z(p, \mathtt{x})\bigr) \le \exp(-\mathtt{x}), \qquad I\!\!P\bigl(\|\boldsymbol{\gamma}\| \le z_1(p, \mathtt{x})\bigr) \le \exp(-\mathtt{x}),$$

where

$$z^2(p, \mathtt{x}) \stackrel{\text{def}}{=} p + \sqrt{6.6 p \mathtt{x}} \vee (6.6 \mathtt{x}), \qquad z_1^2(p, \mathtt{x}) \stackrel{\text{def}}{=} p - 2\sqrt{p \mathtt{x}}.$$

The next important step in our analysis is to check that $\boldsymbol{\vartheta}$ concentrates in a small vicinity $\Theta_0 = \Theta_0(\mathtt{r}_0)$ of the central point $\boldsymbol{\theta}^*$ with a properly selected $\mathtt{r}_0$. The concentration properties of the posterior will be described by using the random quantity

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta \setminus \Theta_0} \exp\bigl\{L(\boldsymbol{\theta})\bigr\} d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\bigl\{L(\boldsymbol{\theta})\bigr\} d\boldsymbol{\theta}} .$$

Obviously $\mathbb{P}\bigl\{\boldsymbol{\vartheta} \notin \Theta_0(\mathtt{r}_0) \bigm| \boldsymbol{Y}\bigr\} \le \rho(\mathtt{r}_0)$. Therefore, small values of $\rho(\mathtt{r}_0)$ indicate a small posterior probability of the set $\Theta \setminus \Theta_0$. The proof only uses condition $(\mathcal{L})$ and the fact that there exists a random set $\Omega(\mathtt{x})$ of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ such that

$$\bigl|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^{\top} D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le \mathtt{r}\, \varrho(\mathtt{r}, \mathtt{x}) \tag{9}$$

for $\mathtt{r} = \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$; cf. the proof of Theorem 19.
Let $\mathtt{b}_0 = \mathtt{b}(\mathtt{r}_0)$ and for the sequence $\mathtt{b}_k = 2^{-k}\mathtt{b}_0$, the radii $\mathtt{r}_0 < \mathtt{r}_1 < \ldots$ be defined by the condition $\mathtt{b}(\mathtt{r}) \ge \mathtt{b}_k > 0$ for $\mathtt{r}_k \le \mathtt{r} < \mathtt{r}_{k+1}$ for all $k \ge 0$ with $\mathtt{b}(\mathtt{r})$ from $(\mathcal{L})$.

Theorem

Suppose the conditions $(\mathcal{L})$, (ED_0), *and* (ED_2). *If* $\mathtt{b}(\mathtt{r})$ *from* $(\mathcal{L})$ *satisfies*

$$\mathtt{r}^2\mathtt{b}^2(\mathtt{r}) \geq \mathtt{x} + 2p + 4z^2(B,\mathtt{x}) + 8\mathtt{r}\,\mathtt{b}(\mathtt{r})\varrho(\mathtt{r},\mathtt{x}), \qquad \mathtt{r} \geq \mathtt{r}_0, \tag{10}$$

then it holds on a set $\Omega(\mathtt{x})$ *of probability at least* $1 - 4\mathrm{e}^{-\mathtt{x}}$

$$\rho(\mathtt{r}_0) \stackrel{\mathrm{def}}{=} \frac{\int_{\Theta\setminus\Theta_0} \exp\bigl\{L(\boldsymbol{\theta})\bigr\}d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\bigl\{L(\boldsymbol{\theta})\bigr\}d\boldsymbol{\theta}} \leq 2\exp\{-\mathtt{x} + \Delta^{+}(\mathtt{r}_0,\mathtt{x})\} \tag{11}$$

with $\Delta^{+}(\mathtt{r}_0,\mathtt{x}) \leq 2\Delta(\mathtt{r}_0,\mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}}$.

Suppose that $\mathtt{b}_0 = \mathtt{b}(\mathtt{r}_0)$ is close to one and $\varrho(\mathtt{r},\mathtt{x})$ small. Condition (10) requires that

$$\mathtt{r}_0^2 > 4z^2(B,\mathtt{x}) + 2p + \mathtt{x}$$

and the value $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows with $\mathtt{r}$.

Use the decomposition

$$L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*) + \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*).$$

$$= I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top.$$

Condition $(\mathcal{L})$ for the expected negative log-likelihood implies

$$-I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \left| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right|^2 \mathtt{b}_k / 2$$

for each $k \geq 0$ and any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k)$. The bound (9) implies on $\Omega(\mathtt{x})$

$$\left| \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right| \leq \mathtt{r}_{k+1}\, \varrho(\mathtt{r}_{k+1}, \mathtt{x}), \qquad \boldsymbol{\theta} \in \Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k),$$

Represent

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta \setminus \Theta_0} \exp\bigl\{ L(\boldsymbol{\theta}) \bigr\} d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\bigl\{ L(\boldsymbol{\theta}) \bigr\} d\boldsymbol{\theta}} = \frac{\exp\bigl\{ m(\boldsymbol{\xi}) \bigr\} \int_{\Theta \setminus \Theta_0} \exp\bigl\{ L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\} d\boldsymbol{\theta}}{\exp\bigl\{ m(\boldsymbol{\xi}) \bigr\} \int_{\Theta_0} \exp\bigl\{ L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr\} d\boldsymbol{\theta}}.$$

By the change of variables $\boldsymbol{\gamma} = D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$, it follows for each k

$$\exp\{m(\boldsymbol{\xi})\} \int_{\Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}$$

$$\leq \exp\{\mathtt{r}_{k+1}\, \varrho(\mathtt{r}_{k+1}, \mathtt{x}) - \|\boldsymbol{\xi}\|^2/2\} \frac{1}{(2\pi)^{p/2}} \int_{\|\boldsymbol{\gamma}\| \geq \mathtt{r}_k} \exp\{-\frac{\mathtt{b}_k \|\boldsymbol{\gamma}\|^2}{2} + \boldsymbol{\xi}^\top \boldsymbol{\gamma}\} d\boldsymbol{\gamma} .$$

Next,

$$\frac{1}{(2\pi)^{p/2}} \int_{\|\boldsymbol{\gamma}\| \geq \mathtt{r}_k} \exp\Bigl(-\frac{\mathtt{b}_k \|\boldsymbol{\gamma}\|^2}{2} + \boldsymbol{\xi}^\top \boldsymbol{\gamma}\Bigr) d\boldsymbol{\gamma}$$

$$\leq \mathtt{b}_k^{-p/2} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{2\mathtt{b}_k}\Bigr) I\!P^{\circ}\bigl(\|\boldsymbol{\gamma} + \mathtt{b}_k^{-1/2} \boldsymbol{\xi}\| \geq \mathtt{b}_k^{1/2} \mathtt{r}_k\bigr)$$

$$\leq \mathtt{b}_k^{-p/2} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{\mathtt{b}_k} - \frac{1}{4} \mathtt{b}_k \mathtt{r}_k^2 + \frac{p}{2}\Bigr). \tag{12}$$

Here we have used the bound for a standard normal vector $\boldsymbol{\gamma}$ and $\boldsymbol{u} = \mathtt{b}_k^{-1/2} \boldsymbol{\xi} \in I\!R^p$. (8) and (12) imply (11).

Now the bound $\|\boldsymbol{\xi}\| \leq z(B, \mathtt{x})$ holding with a dominating probability and (10) imply

$$\sum_{k=0}^{\infty} \exp\{m(\boldsymbol{\xi})\} \int_{\Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} d\boldsymbol{\theta}$$

$$\leq \sum_{k=0}^{\infty} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{\mathtt{b}_k} - \frac{1}{4} \mathtt{b}_k \mathtt{r}_k^2 + \frac{p}{2} \log\bigl(e/\mathtt{b}_k\bigr) + \mathtt{r}_{k+1}\, \varrho(\mathtt{r}_{k+1}, \mathtt{x}) \Bigr)$$

$$\leq \sum_{k=0}^{\infty} \exp(-\mathtt{x}/\mathtt{b}_k) \leq 2\mathrm{e}^{-\mathtt{x}}$$

and (11) follows in view of $\mathtt{b} \log(e/\mathtt{b}) \leq 1$ for $\mathtt{b} \leq 1$.

Theorem

Suppose (2) *for* $\mathtt{r} = \mathtt{r}_0$ *and* (11). *Then for any nonnegative function* $f(\cdot)$ *on* $I\!R^p$ *, it holds on* $\Omega(\mathtt{x})$

$$I\!E^{\circ}\bigl\{f\bigl(D_0(\boldsymbol{\vartheta}-\breve{\boldsymbol{\theta}})\bigr)\,\mathbb{1}_{\mathtt{r}_0}\bigr\} \geq \exp\bigl\{-\Delta^{-}(\mathtt{r}_0,\mathtt{x})\bigr\}\,I\!E\Bigl\{f(\boldsymbol{\gamma})\,\mathbb{1}\bigl(\|\boldsymbol{\gamma}+\boldsymbol{\xi}\|\leq\mathtt{r}_0\bigr)\Bigr\},$$

where

$$\Delta^{-}(\mathtt{r}_0,\mathtt{x}) = \Delta^{+}(\mathtt{r}_0,\mathtt{x}) + \rho(\mathtt{r}_0).$$

Lower bound. Proof

On the set $\Omega(\mathtt{x})$, it holds by (6) with $f(\cdot) = 1$:

$$\begin{aligned}
\int \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\}\, d\boldsymbol{\theta} &\le \int_{\Theta_0} \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\}\, d\boldsymbol{\theta} + \int_{\Theta\setminus\Theta_0} \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\}\, d\boldsymbol{\theta} \\
&\le \bigl\{1+\rho(\mathtt{r}_0)\bigr\} \int_{\Theta_0} \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\}\, d\boldsymbol{\theta} \\
&\le \bigl\{1+\rho(\mathtt{r}_0)\bigr\} \exp\bigl\{\Delta(\mathtt{r}_0,\mathtt{x}) - m(\boldsymbol{\xi}) + \nu(\mathtt{r}_0)\bigr\} \\
&\le \exp\bigl\{\Delta(\mathtt{r}_0,\mathtt{x}) - m(\boldsymbol{\xi}) + \nu(\mathtt{r}_0) + \rho(\mathtt{r}_0)\bigr\}.
\end{aligned}$$

This and the bound (7) imply

$$\begin{aligned}
&\frac{\exp\bigl\{m(\boldsymbol{\xi})\bigr\} \int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\} f\bigl(D_0(\boldsymbol{\theta}-\breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta}}{\exp\bigl\{m(\boldsymbol{\xi})\bigr\} \int \exp\bigl\{L(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr\}\, d\boldsymbol{\theta}} \\
&\quad \ge \frac{\exp\bigl\{-\Delta(\mathtt{r}_0,\mathtt{x})\bigr\} \int \phi(\boldsymbol{u}) f(\boldsymbol{u})\, \mathbb{1}\bigl\{\|\boldsymbol{u}+\boldsymbol{\xi}\| \le \mathtt{r}_0\bigr\} d\boldsymbol{u}}{\exp\bigl\{\Delta(\mathtt{r}_0,\mathtt{x}) + \nu(\mathtt{r}_0) + \rho(\mathtt{r}_0)\bigr\}} \\
&\quad \ge \exp\bigl\{-\Delta^{-}(\mathtt{r}_0,\mathtt{x})\bigr\}\, \mathbb{E}\bigl[f(\boldsymbol{\gamma})\, \mathbb{1}\bigl\{\|\boldsymbol{\gamma}+\boldsymbol{\xi}\| \le \mathtt{r}_0\bigr\}\bigr].
\end{aligned}$$

Define $\mathcal{C}^{\circ}(A) = \{\boldsymbol{\theta} \colon D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}}) \in A\}$. Then

$$I\!P\bigl(\mathcal{C}^{\circ}(A) \,\big|\, \boldsymbol{Y}\bigr) \approx I\!P(\boldsymbol{\gamma} \in A) \pm \mathtt{C}\Delta(\mathtt{r}_0, \mathtt{x}).$$

Unfortunately, the quantities $\breve{\boldsymbol{\theta}}$ and D_0^2 are unknown and cannot be used for building the elliptic credible sets.

A natural question: empirical counterparts.

Theorem

Let a vector $\widehat{\boldsymbol{\theta}}$ *and a symmetric matrix* $\widehat{D}$ *fulfill*

$$\|D_0(\widehat{\boldsymbol{\theta}} - \breve{\boldsymbol{\theta}})\| \le \beta, \qquad \widehat{D}^2 \le a^2 D_0^2, \qquad \operatorname{tr}\bigl(D_0^{-1}\widehat{D}^2 D_0^{-1} - I_p\bigr)^2 \le \epsilon^2.$$

Then with $\tau = \frac{1}{2}\sqrt{a^2\beta^2 + \epsilon^2}$, *it holds on a random set* $\Omega(\mathtt{x})$ *of probability* $1 - 5\mathrm{e}^{-\mathtt{x}}$

$$I\!P\bigl(\widehat{D}(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\theta}}) \in A \,\big|\, \boldsymbol{Y}\bigr) \ge \exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 3\mathrm{e}^{-\mathtt{x}}\bigr)\bigl\{I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - \tau\bigr\} - \mathrm{e}^{-\mathtt{x}},$$

$$I\!P\bigl(\widehat{D}(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\theta}}) \in A \,\big|\, \boldsymbol{Y}\bigr) \le \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}}\bigr)\bigl\{I\!P\bigl(\boldsymbol{\gamma} \in A\bigr) + \tau\bigr\} + \mathrm{e}^{-\mathtt{x}}.$$

Credible sets. Cont.

Denote $U = \widehat{D} D_0^{-1}$ and $\boldsymbol{\eta} = D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}})$, and $\boldsymbol{\beta} = D_0(\widehat{\boldsymbol{\theta}} - \breve{\boldsymbol{\theta}})$. Then

$$I\!\!P\bigl(\widehat{D}(\boldsymbol{\vartheta} - \widehat{\boldsymbol{\theta}}) \in A \,\big|\, \boldsymbol{Y}\bigr) = I\!\!P\bigl(U(\boldsymbol{\eta} - \boldsymbol{\beta}) \in A \,\big|\, \boldsymbol{Y}\bigr) \approx I\!\!P\bigl(U(\boldsymbol{\gamma} - \boldsymbol{\beta}) \in A \,\big|\, \boldsymbol{Y}\bigr).$$

Now the result follows from Theorem 1 and

Lemma

Let $I\!\!P_0 = \mathcal{N}(0, I_p)$ *and* $I\!\!P_1 = \mathcal{N}(\boldsymbol{\beta}, (U^\top U)^{-1})$ *some non-degenerated matrix* U. *If*

$$\|U^\top U - I_p\|_\infty \le \boldsymbol{\epsilon} \le 1/2,$$

then $\mathcal{K}(I\!\!P_0, I\!\!P_1) = -I\!\!E_0 \log \frac{dI\!\!P_1}{dI\!\!P_0}$ *fulfills*

$$2\mathcal{K}(I\!\!P_0, I\!\!P_1) \le \operatorname{tr}(U^\top U - I_p)^2 + (1+\boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2 \le \boldsymbol{\epsilon}^2 p + (1+\boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2.$$

For any measurable set $A \subset I\!\!R^p$, *it holds with* $\boldsymbol{\gamma} \sim \mathcal{N}(0, I_p)$

$$\bigl|I\!\!P_0(A) - I\!\!P_1(A)\bigr| = \bigl|I\!\!P\bigl(\boldsymbol{\gamma} \in A\bigr) - I\!\!P\bigl(U(\boldsymbol{\gamma} - \boldsymbol{\beta}) \in A\bigr)\bigr| \le \sqrt{\mathcal{K}(I\!\!P_0, I\!\!P_1)/2}.$$

Proof

It holds

$$2 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0}(\boldsymbol{\gamma}) = \log \det(U^\top U) - (\boldsymbol{\gamma} - \boldsymbol{\beta})^\top U^\top U(\boldsymbol{\gamma} - \boldsymbol{\beta}) + \|\boldsymbol{\gamma}\|^2$$

with $\boldsymbol{\gamma}$ standard normal and

$$2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) = -2\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0} = -\log \det(U^\top U) + \operatorname{tr}(U^\top U - I_p) + \boldsymbol{\beta}^\top U^\top U \boldsymbol{\beta}.$$

Let a_j be the j th eigenvalue of $U^\top U - I_p$. $\|U^\top U - I_p\|_\infty \le \boldsymbol{\epsilon} \le 1/2$ yields $|a_j| \le 1/2$ and

$$\begin{aligned} 2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) &= \boldsymbol{\beta}^\top U^\top U \boldsymbol{\beta} + \sum_{j=1}^{p} \{a_j - \log(1 + a_j)\} \le (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} a_j^2 \\ &\le (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2 + \operatorname{tr}(U^\top U - I_p)^2 \le (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2 + \boldsymbol{\epsilon}^2 p. \end{aligned}$$

This implies by Pinsker's inequality

$$\sup_A |\mathbb{P}_0(A) - \mathbb{P}_1(A)| \le \sqrt{\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)/2}.$$

Credible set as frequentist CS

Define

$$\mathcal{C}(A_\alpha) = \bigl\{\boldsymbol{\theta} \colon \widehat{D}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}) \in A_\alpha\bigr\},$$

where $\widehat{D}^2 \approx D_0^2$ and $\widehat{\boldsymbol{\theta}} \approx \breve{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + D_0^{-1}\boldsymbol{\xi} \approx \widetilde{\boldsymbol{\theta}}$. Then

$$I\!P^{\circ}\bigl\{\mathcal{C}(A_\alpha)\bigr\} \approx I\!P(\boldsymbol{\gamma} \in A_\alpha) \pm \mathtt{C}\Delta(\mathtt{r}_0, \mathtt{x}).$$

$\mathcal{C}(A_\alpha)$ is completely data-based, can be constructed by Bayesian simulations and $I\!P^{\circ}\bigl\{\mathcal{C}(A_\alpha)\bigr\} \approx \alpha$!

Question: can one use $\mathcal{C}(A_\alpha)$ as a frequentist confidence set?

The construction of $\mathcal{C}(A_\alpha)$ perfectly matches the usual frequentist asymptotic CS.

- Under PA $\mathcal{C}(A_\alpha)$ is an asymptotic α -CS.
- If PA-PW, the CS $\mathcal{C}(A_\alpha)$ can be totally wrong, cf. [Cox, 1993] or [Kleijn and van der Vaart, 2012].

Outline

1 Bernstein – von Mises Theorem
- BvM Theorem
- Local Gaussian approximation of the posterior
- Tail posterior probability and contraction
- Credible sets

2 Semiparametric estimation
- Motivation
- Linear models
- General semiparametric setup

3 Penalized MLE and effective dimension
- Curse of dimension
- Effective dimension
- Fisher and Wilks expansions
- Concentration and large deviations
- A bound for the norm of a vector stochastic process

4 Confidence estimation using bootstrap
- Likelihood-based confidence set
- Multipier bootstrap
- Conditions

Data $\boldsymbol{Y}$ with DGP $\boldsymbol{Y} \sim I\!\!P$.

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

$\boldsymbol{\theta}$, target, $\dim(\boldsymbol{\theta}) = p$, $\quad \boldsymbol{\eta}$, nuisance, $\dim(\boldsymbol{\eta}) = q$, $\quad p^{*} = p + q$.

Goal: inference on $\boldsymbol{\theta}$.

Examples in mind:

- an inverse problem with error in operator;
  $\boldsymbol{Y} = A\boldsymbol{\theta} + \boldsymbol{\varepsilon}$, observed $\boldsymbol{Y}$ and $\widehat{A}$, operator A as nuisance;
- transformation models $\Lambda Y = f(X) + \epsilon$: the transfer Λ or regression function $\boldsymbol{f}$ as nuisance;
- Hidden Markov Chains $Y_t \sim P_{f(X_t,\boldsymbol{\theta})}$: the whole hidden path X_t as nuisance.
- Error-in-variable regression $Y_i = f(X_i) + \epsilon_i$, $Z_i = X_i + \xi_i$: the whole unobserved design $\boldsymbol{X}$ as nuisance.

Profile MLE (pMLE)

SPA : $\boldsymbol{Y} \sim \mathbb{P} \in \left(\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}},\ \boldsymbol{\theta} \in \Theta, \boldsymbol{\eta} \in H\right)$

Log-likelihood: $L(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{d\mathbb{P}_{\boldsymbol{\theta},\boldsymbol{\eta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$

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

Murphy, van der Vaart (2000), Kosorok (2005, 2008): Under PA $\mathbb{P} = \mathbb{P}_{\boldsymbol{\theta}^*,\boldsymbol{\eta}^*}$, the pMLE $\widetilde{\boldsymbol{\theta}}$ is

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

Limitations:

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

$(\boldsymbol{\theta}, \boldsymbol{\eta})$ -setup:

$$\boldsymbol{Y} = \Psi^{\top} \boldsymbol{\theta}^{*} + \Phi^{\top} \boldsymbol{\eta}^{*} + \boldsymbol{\varepsilon},$$

where Ψ is $p \times n$ matrix of essential factors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_p$, Φ is $q \times n$ -matrix of nuisance factors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_q$.

$\boldsymbol{\upsilon}$ -setup:

$$\boldsymbol{Y} = \Upsilon^{\top} \boldsymbol{\upsilon}^{*} + \boldsymbol{\varepsilon}$$

with p^{*} factors $(\boldsymbol{\psi}_j), (\boldsymbol{\phi}_m)$, and the target of estimation is a linear mapping $\boldsymbol{\theta}^{*} = P \boldsymbol{\upsilon}^{*}$ for a given projector $P : I\!R^{p^{*}} \to I\!R^{p}$.

$\boldsymbol{\upsilon}$ -setup:

$$\boldsymbol{Y} = \varUpsilon^{\top}\boldsymbol{\upsilon}^{*} + \boldsymbol{\varepsilon} = \varPsi^{\top}\boldsymbol{\theta}^{*} + \varPhi^{\top}\boldsymbol{\eta}^{*} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}\boldsymbol{\varepsilon} = 0, \operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^{2} I_{n}.$$

Target: $\boldsymbol{\theta}^{*} = P\boldsymbol{\upsilon}^{*}$.

Profile qMLE 1: $\widetilde{\boldsymbol{\theta}} = P\widetilde{\boldsymbol{\upsilon}} = P\bigl(\varUpsilon\varUpsilon^{\top}\bigr)^{-1}\varUpsilon\boldsymbol{Y} = S\boldsymbol{Y}$, $\quad S = P\bigl(\varUpsilon\varUpsilon^{\top}\bigr)^{-1}\varUpsilon$.

Profile qMLE 2: $\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} \breve{L}(\boldsymbol{\theta})$, $\quad \breve{L}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sup_{\boldsymbol{\upsilon}:\, P\boldsymbol{\upsilon}=\boldsymbol{\theta}} L(\boldsymbol{\upsilon})$.

Theorem

$$\begin{aligned} \mathbb{E}\widetilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^{*}, \\ \operatorname{Var}(\widetilde{\boldsymbol{\theta}}) &= S\operatorname{Var}(\boldsymbol{\varepsilon})S^{\top} = \sigma^{2}SS^{\top} = \sigma^{2}P\bigl(\varUpsilon\varUpsilon^{\top}\bigr)^{-1}P^{\top}. \end{aligned}$$

Model:

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

Theorem

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

$$\widetilde{\boldsymbol{\theta}} = \bigl(\breve{\Psi}\breve{\Psi}^{\top}\bigr)^{-1}\breve{\Psi}\boldsymbol{Y},$$

$$\breve{\Psi} = \Psi - \Psi \Pi_{\boldsymbol{\eta}} = \Psi - \Psi\Phi^{\top}\bigl(\Phi\Phi^{\top}\bigr)^{-1}\Phi.$$

Model:

$$\boldsymbol{Y} = \varUpsilon^\top \boldsymbol{\upsilon}^* + \boldsymbol{\varepsilon} = \varPsi^\top \boldsymbol{\theta}^* + \varPhi^\top \boldsymbol{\eta}^* + \boldsymbol{\varepsilon} \qquad I\!E \boldsymbol{\varepsilon} = 0, \ \operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^2 I_n .$$

Theorem (Gauss-Markov)

1. $\widetilde{\boldsymbol{\theta}} = S\boldsymbol{Y}$ *with* $S = P\bigl(\varUpsilon\varUpsilon^\top\bigr)^{-1}\varUpsilon$ *fulfills*

$$\begin{aligned}
I\!E\widetilde{\boldsymbol{\theta}} &= \boldsymbol{\theta}^* = P\boldsymbol{\upsilon}^*, \\
I\!E\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|^2 &= \operatorname{Var}(\widetilde{\boldsymbol{\theta}}) = \sigma^2 P\bigl(\varUpsilon\varUpsilon^\top\bigr)^{-1}P^\top = \sigma^2\bigl(\breve{\varPsi}\breve{\varPsi}^\top\bigr)^{-1}, \\
\breve{\varPsi} &= \varPsi - \varPsi \varPi_{\boldsymbol{\eta}} \\
\varPi_{\boldsymbol{\eta}} &= \varPhi^\top\bigl(\varPhi\varPhi^\top\bigr)^{-1}\varPhi.
\end{aligned}$$

2. This risk is minimal in the class of all unbiased linear estimates of $\boldsymbol{\theta}^*$ *.*

Model:

$$\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta}^{*} + \Phi^{\top}\boldsymbol{\eta}^{*} + \boldsymbol{\varepsilon} \qquad \mathbb{E}\boldsymbol{\varepsilon} = 0, \; \operatorname{Var}(\boldsymbol{\varepsilon}) = \sigma^{2} I_{n}.$$

Define

$$\breve{D}_0^2 = \sigma^{-2}\breve{\Psi}\breve{\Psi}^{\top}, \qquad \breve{\Psi} = \Psi - \Psi \Pi_{\boldsymbol{\eta}}.$$

Theorem

Let the matrix $\breve{D}_0^2$ *be non-degenerated. It holds*

$$\begin{aligned} 2\bigl\{\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^{*})\bigr\} &= \bigl\|\breve{D}_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr)\bigr\|^{2} = \|\breve{\boldsymbol{\xi}}\|^{2}, \\ \breve{\boldsymbol{\xi}} &= \breve{D}_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr), \qquad \mathbb{E}\breve{\boldsymbol{\xi}} = 0, \; \operatorname{Var}(\breve{\boldsymbol{\xi}}) = I_{p}\,. \end{aligned}$$

If $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^{2} I_{n})$ *, then* $\breve{\boldsymbol{\xi}}$ *is standard normal in* $\mathbb{R}^{p}$ *and*

$$2\bigl\{\breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^{*})\bigr\} \sim \chi_{p}^{2}\,.$$

SPA

$$\boldsymbol{Y} \sim I\!P \in \bigl(I\!P_{\boldsymbol{\theta},\boldsymbol{\eta}},\, \boldsymbol{\theta} \in \Theta, \boldsymbol{\eta} \in H\bigr)$$

Log-likelihood:

$$L(\boldsymbol{\theta},\boldsymbol{\eta}) = \frac{dI\!P_{\boldsymbol{\theta},\boldsymbol{\eta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$$

Profile MLE:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} L(\boldsymbol{\theta},\boldsymbol{\eta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \breve{L}(\boldsymbol{\theta}),$$

$$\breve{L}(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}} L(\boldsymbol{\theta},\boldsymbol{\eta}).$$

$\boldsymbol{\upsilon}$ -setup: $\boldsymbol{\upsilon} = (\boldsymbol{\theta},\boldsymbol{\eta})\,,\; L(\boldsymbol{\upsilon}) = L(\boldsymbol{\theta},\boldsymbol{\eta})\,,$

$$\widetilde{\boldsymbol{\upsilon}} = \operatorname*{argmax}_{\boldsymbol{\upsilon}} L(\boldsymbol{\upsilon}), \qquad \widetilde{\boldsymbol{\theta}} = P\widetilde{\boldsymbol{\upsilon}}$$

Full dimensional expansion. Main definitions

For $L(\boldsymbol{\upsilon}) = L(\boldsymbol{\theta}, \boldsymbol{\eta})$, define

$$\boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \varUpsilon} I\!E L(\boldsymbol{\upsilon}),$$

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}} \max_{\boldsymbol{\eta}} I\!E L(\boldsymbol{\theta}, \boldsymbol{\eta}) = P \boldsymbol{\upsilon}^* .$$

Also

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 I\!E L(\boldsymbol{\upsilon}^*),$$

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} \mathcal{D}_0^{-1} \nabla L(\boldsymbol{\upsilon}^*),$$

$$\mathcal{V}_0^2 = \operatorname{Var}\{\nabla L(\boldsymbol{\upsilon}^*)\} \quad (= \mathcal{D}_0^2 \quad \text{under PA})$$

and

$$\varUpsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} \colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}.$$

- Concentration and large deviations: fix $\mathtt{r}_0$ ensuring

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\upsilon}} \notin \Upsilon_{\circ}(\mathtt{r}_0)\bigr) \le \mathrm{e}^{-\mathtt{x}},$$

  where $\Upsilon_{\circ}(\mathtt{r}) \stackrel{\text{def}}{=} \bigl\{\boldsymbol{\theta}\colon \|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\| \le \mathtt{r}\bigr\}$.

- Local quadratic approximation of the expected log-likelihood:

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})} \frac{2\mathbb{E}L(\boldsymbol{\upsilon}^{*}) - 2\mathbb{E}L(\boldsymbol{\upsilon})}{\|\mathcal{D}_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^{*})\|^2} \le \delta(\mathtt{r}).$$

- Local linear approximation of the stochastic component: on $\Omega(\mathtt{x})$, for $\zeta(\boldsymbol{\upsilon}) \stackrel{\text{def}}{=} L(\boldsymbol{\upsilon}) - \mathbb{E}L(\boldsymbol{\upsilon})$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})} \bigl|\mathcal{D}_0^{-1}\bigl\{\nabla\zeta(\boldsymbol{\upsilon}) - \nabla\zeta(\boldsymbol{\upsilon}^{*})\bigr\}\bigr| \le \varrho(\mathtt{r}, \mathtt{x}).$$

- Overall error of the Fisher expansion $\mathtt{r}_0\bigl\{\delta(\mathtt{r}_0) + \varrho(\mathtt{r}_0, \mathtt{x})\bigr\}$, of the Wilks $\mathtt{r}_0^2\bigl\{\delta(\mathtt{r}_0) + \varrho(\mathtt{r}_0, \mathtt{x})\bigr\}$.

$$\tilde{\boldsymbol{\upsilon}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} L(\boldsymbol{\upsilon}), \qquad \boldsymbol{\upsilon}^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\upsilon} \in \Upsilon} \mathbb{E} L(\boldsymbol{\upsilon}),$$

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\upsilon}^*), \qquad \boldsymbol{\xi} \stackrel{\text{def}}{=} \mathcal{D}_0^{-1} \nabla L(\boldsymbol{\upsilon}^*).$$

Theorem

On a set $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| \mathcal{D}_0 (\tilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \boldsymbol{\xi} \bigr\| \leq \diamondsuit(\mathtt{r}_0, \mathtt{x}),$$

$$\Bigl| L(\tilde{\boldsymbol{\upsilon}}) - L(\boldsymbol{\upsilon}^*) - \frac{\|\boldsymbol{\xi}\|^2}{2} \Bigr| \leq \Delta(\mathtt{r}_0, \mathtt{x}).$$

Here $\diamondsuit(\mathtt{r}_0, \mathtt{x})$ *and* $\Delta(\mathtt{r}_0, \mathtt{x})$ *are explicit error terms.*
The vector $\boldsymbol{\xi}$ *fulfills*

$$\mathbb{P}\bigl(\|\boldsymbol{\xi}\| \geq z(B, \mathtt{x})\bigr) \leq 2\mathrm{e}^{-\mathtt{x}},$$

where $B = \operatorname{Var}(\boldsymbol{\xi}) = \mathcal{D}_0^{-1} \mathcal{V}_0^2 \mathcal{D}_0^{-1}$ *, so that* $z^2(B, \mathtt{x}) \asymp p^* + \mathtt{x}$ *.*

Problems: the value of $\|\boldsymbol{\xi}\|^2$ is of order of the full dimension p^*.

Corollaries for $\widetilde{\boldsymbol{\theta}} = P\widetilde{\boldsymbol{\upsilon}}$?

Consider the block representation:

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A \\ A^\top & H_0^2 \end{pmatrix}, \qquad \nabla = \nabla L(\boldsymbol{\upsilon}^*) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) \\ \nabla_{\boldsymbol{\eta}} L(\boldsymbol{\theta}^*, \boldsymbol{\eta}^*) \end{pmatrix} = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix},$$

Define $\breve{D}_0^{-2}$ as the left upper block of $\mathcal{D}_0^{-2}$:

$$\breve{D}_0^2 = D_0^2 - A H_0^{-2} A^\top$$

and

$$\breve{\boldsymbol{\xi}} \stackrel{\mathrm{def}}{=} \breve{D}_0^{-1} \bigl(\nabla_{\boldsymbol{\theta}} - A H_0^{-2} \nabla_{\boldsymbol{\eta}} \bigr)$$

Inference on θ^*

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A \\ A^\top & H_0^2 \end{pmatrix}, \qquad \nabla = \nabla L(\boldsymbol{\upsilon}^*) = \begin{pmatrix} \nabla_{\boldsymbol{\theta}} \\ \nabla_{\boldsymbol{\eta}} \end{pmatrix},$$

$$\breve{D}_0^2 = D_0^2 - A H_0^{-2} A^\top \qquad \breve{\boldsymbol{\xi}} \stackrel{\text{def}}{=} \breve{D}_0^{-1} \breve{\nabla}_{\boldsymbol{\theta}} = \breve{D}_0^{-1} \bigl(\nabla_{\boldsymbol{\theta}} - A H_0^{-2} \nabla_{\boldsymbol{\eta}} \bigr)$$

Theorem

On a set $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| \breve{D}_0 (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}} \bigr\| \leq \diamondsuit(\mathtt{r}_0, \mathtt{x}),$$

$$\Bigl| \breve{L}(\widetilde{\boldsymbol{\theta}}) - \breve{L}(\boldsymbol{\theta}^*) - \frac{\|\breve{\boldsymbol{\xi}}\|^2}{2} \Bigr| \leq \Delta(\mathtt{r}_0, \mathtt{x}) \leq \mathtt{C}\, p \, \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

Here $\diamondsuit(\mathtt{x})$ *and* $\Delta(\mathtt{x})$ *are explicit error terms. The vector* $\breve{\boldsymbol{\xi}}$ *fulfills*

$$\mathbb{P}(\|\breve{\boldsymbol{\xi}}\| \geq z(\breve{B}, \mathtt{x})) \leq 2\mathrm{e}^{-\mathtt{x}},$$

where $\breve{B} = \operatorname{Var}(\breve{\boldsymbol{\xi}}) = \breve{D}_0^{-1} \operatorname{Var}(\breve{\nabla}) \breve{D}_0^{-1}$ *, so that* $z^2(\breve{B}, \mathtt{x}) \asymp p + \mathtt{x}$ *.*

Concentration and large deviation for the PMLE $\widetilde{\theta}$

Steps:

- Concentration of $\widetilde{\boldsymbol{\upsilon}}$ on $\Upsilon_{\circ}(\mathtt{r}_0)$ for $\mathtt{r}_0^2 \asymp p^* + \mathtt{x}$;
- Full dimensional Fisher expansion: on $\Omega(\mathtt{x})$

$$\bigl\| \mathcal{D}_0(\widetilde{\boldsymbol{\upsilon}} - \boldsymbol{\upsilon}^*) - \boldsymbol{\xi} \bigr\| \leq \diamondsuit(\mathtt{r}_0, \mathtt{x});$$

- Fisher expansion for $\widetilde{\boldsymbol{\theta}}$: on $\Omega(\mathtt{x})$

$$\bigl\| \breve{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \breve{\boldsymbol{\xi}} \bigr\| \leq \diamondsuit(\mathtt{r}_0, \mathtt{x});$$

- A deviation bound

$$I\!P\bigl(\|\breve{\boldsymbol{\xi}}\| \geq z(\breve{B}, \mathtt{x}) \bigr) \leq 2\mathrm{e}^{-\mathtt{x}}$$

Imply concentration of $\widetilde{\boldsymbol{\theta}}$ on $\Theta_0(\breve{\mathtt{r}}_0)$ for $\breve{\mathtt{r}}_0 = z(\breve{B}, \mathtt{x}) + \diamondsuit(\mathtt{r}_0, \mathtt{x})$:

$$I\!P\Bigl\{ \bigl\| \breve{D}_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \bigr\| \geq z(\breve{B}, \mathtt{x}) + \diamondsuit(\mathtt{r}_0, \mathtt{x}) \Bigr\} \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Outline

1 Bernstein – von Mises Theorem
- BvM Theorem
- Local Gaussian approximation of the posterior
- Tail posterior probability and contraction
- Credible sets

2 Semiparametric estimation
- Motivation
- Linear models
- General semiparametric setup

3 Penalized MLE and effective dimension
- Curse of dimension
- Effective dimension
- Fisher and Wilks expansions
- Concentration and large deviations
- A bound for the norm of a vector stochastic process

4 Confidence estimation using bootstrap
- Likelihood-based confidence set
- Multipier bootstrap
- Conditions

Let $p = p_n \to \infty$. We know

$$\diamondsuit_n(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p_n + \mathtt{x})^2}{n}}, \qquad \Delta_n(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p_n + \mathtt{x})^3}{n}}, \qquad \|\boldsymbol{\xi}_n\|^2 \le p_n + \mathtt{C}\mathtt{x}.$$

- $p_n/n \to 0$: Consistency:

$$\|\sqrt{\mathbb{F}_{\boldsymbol{\theta}^*}}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\| = n^{-1/2}\{\|\boldsymbol{\xi}_n\| \pm \diamondsuit_n(\mathtt{x})\} \le \sqrt{\frac{p_n + \mathtt{C}\mathtt{x}}{n}} \pm \mathtt{C}\,\frac{p_n + \mathtt{x}}{n}$$

- $p_n^2/n \to 0$ – Fisher expansion, root-n normality;

$$\sqrt{n\mathbb{F}_{\boldsymbol{\theta}^*}}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \boldsymbol{\xi}_n \pm \diamondsuit_n(\mathtt{x}), \qquad \text{expansion of the MLE}$$

$$\sqrt{2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} = \|\boldsymbol{\xi}_n\| \pm 3\diamondsuit_n(\mathtt{x}), \qquad \text{square-root excess}$$

$$p_n^{-1/2} L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = p_n^{-1/2}\|\boldsymbol{\xi}_n\|^2/2 \pm \mathtt{C}\diamondsuit_n(\mathtt{x}), \qquad \text{likelihood ratio tests, model selection}$$

- $p_n^3/n \to 0$ – Wilks approximation, BvM Theorem.

Let $\operatorname{pen}(\boldsymbol{\theta})$ be a penalty function on Θ .

Large $\operatorname{pen}(\boldsymbol{\theta}) \iff$ rough $\boldsymbol{\theta}$.

Small $\operatorname{pen}(\boldsymbol{\theta}) \iff$ smooth $\boldsymbol{\theta}$.

Structural assumption – the true value $\boldsymbol{\theta}^*$ is smooth – $\operatorname{pen}(\boldsymbol{\theta}_0)$ is (relatively) small.

A penalized (quasi) MLE approach leads to maximizing the penalized log-likelihood:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \bigl\{ L(\boldsymbol{\theta}) - \operatorname{pen}(\boldsymbol{\theta}) \bigr\}.$$

New target:

$$\boldsymbol{\theta}^*_{\operatorname{pen}} = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \bigl\{ \mathbb{E} L(\boldsymbol{\theta}) - \operatorname{pen}(\boldsymbol{\theta}) \bigr\}.$$

In general, $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^*_{\operatorname{pen}}$: “modeling bias” issue.

Quadratic penalization

Important special case – a quadratic penalty $\operatorname{pen}(\boldsymbol{\theta}) = \|G\boldsymbol{\theta}\|^2/2$ for a given symmetric matrix G^2. Denote

$$L_G(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - \|G\boldsymbol{\theta}\|^2/2,$$

$$\widetilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L_G(\boldsymbol{\theta}).$$

The use of a penalty changes the target of estimation which is now defined as

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

In general $\boldsymbol{\theta}^* \neq \boldsymbol{\theta}^*_G$.

The modeling bias can be measured by $\|G\boldsymbol{\theta}^*\|^2$, yielding the “bias-variance” trade-off:

$$\mathbb{E}\|\boldsymbol{\xi}_G\|^2 \asymp \|G\boldsymbol{\theta}^*\|^2$$

Let $V_0^2 = \mathrm{Var}\bigl\{\nabla L(\boldsymbol{\theta}_G^*)\bigr\}$.

Typically V_0^2 measures the variability of the process $L(\cdot)$ and $L_G(\cdot)$.

Let also D_G^2 be a penalized information matrix

$$D_G^2 = -\nabla^2 \mathbb{E} L_G(\boldsymbol{\theta}_G^*) = D_0^2 + G^2$$

with $D_0^2 = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}_G^*)$.

The effective dimension $\mathtt{p}_G$ is defined as the trace of the matrix $B_G \stackrel{\mathrm{def}}{=} D_G^{-1} V_0^2 D_G^{-1}$:

$$\mathtt{p}_G \stackrel{\mathrm{def}}{=} \mathrm{tr}\bigl(B_G\bigr) = \mathbb{E}\|\boldsymbol{\xi}_G\|^2$$

for $\boldsymbol{\xi}_G = D_G^{-1} \nabla L(\boldsymbol{\theta}_G^*)$.

Examples of computing the effective dimension

Let

$$V_0^2 \;=\; D_0^2 = \sigma^2 I_p,$$
$$G^2 \;=\; \operatorname{diag}\{g_1^2 \ge g_2^2 \ge \dots g_p^2\}$$

Then

$$D_G^2 \;=\; D_0^2 + G^2 = \operatorname{diag}\{\sigma^2 + g_1^2, \dots, \sigma^2 + g_p^2\},$$
$$B_G \;=\; \operatorname{diag}\{(1+\sigma^{-2} g_1^2)^{-1}, \dots, (1+\sigma^{-2} g_p^2)^{-1}\}.$$

Block penalization

G is of a block structure: $G = \operatorname{diag}\{0, G_1\}$.

The first block of dimension p_0 corresponds to the unconstrained part of the parameter vector

the second block of dimension p_1 corresponds to the low energy component.

Assume for simplicity that $G_1 = g I_{p_1}$. Then

$$\mathtt{p}_G = \operatorname{tr} B_G = p_0 + p_1 / \bigl(1 + \sigma^{-2} g^2\bigr).$$

The impact of G_1 in the effective dimension is inessential if $g^2/\sigma^2 \gg p_1/p_0$.

A Sobolev smoothness constraint

For $\beta > 1/2$,

$$G^2 = \operatorname{diag}\{g_1^2, \ldots, g_p^2\}$$
$$g_j = Lj^{\beta}$$

The value β is usually considered as the Sobolev smoothness parameter.

It holds

$$\mathtt{p}_G = \sum_{j=1}^{p} \frac{1}{1 + L^2 j^{2\beta}/\sigma^2}\,.$$

Define also the index $\mathtt{p}_e$ as the largest j satisfying $Lj^{\beta} \le \sigma$.

$\beta > 1/2$ yields $\mathtt{p}_G \le \mathtt{C}(\beta)\mathtt{p}_e$ for some constant $\mathtt{C}(\beta)$ depending on β only.

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

Theorem

On a set $\Omega(\mathtt{x})$ *with* $I\!\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}e^{-\mathtt{x}}$

$$\bigl\| D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi} \bigr\| \leq \diamondsuit(\mathtt{x}),$$

$$\Bigl| L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) - \frac{\|\boldsymbol{\xi}\|^2}{2} \Bigr| \leq \Delta(\mathtt{x})$$

with

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

$$\tilde{\boldsymbol{\theta}}_G \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} L_G(\boldsymbol{\theta}), \qquad \boldsymbol{\theta}^*_G \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta}\in\Theta} \mathbb{E} L_G(\boldsymbol{\theta})$$

Theorem

On a set $\Omega(\mathtt{x})$ *with* $\mathbb{P}\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$

$$\bigl\| D_G(\tilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}^*_G) - \boldsymbol{\xi}_G \bigr\| \leq \diamondsuit_G(\mathtt{x}),$$

$$\Bigl| L_G(\tilde{\boldsymbol{\theta}}_G) - L_G(\boldsymbol{\theta}^*_G) - \frac{\|\boldsymbol{\xi}_G\|^2}{2} \Bigr| \leq \Delta_G(\mathtt{x})$$

with

$$D_G^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L_G(\boldsymbol{\theta}^*_G) = -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*_G) + G^2,$$

$$\boldsymbol{\xi}_G \stackrel{\text{def}}{=} D_G^{-1} \nabla L_G(\boldsymbol{\theta}^*).$$

$(\mathcal{L})$ *For each* $\mathtt{r}$ *, there exists* $\mathtt{b}(\mathtt{r}) > 0$ *such that* $\mathtt{r}\mathtt{b}(\mathtt{r}) \to \infty$ *as* $\mathtt{r} \to \infty$ *and*

$$\frac{-2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \geq \mathtt{b}(\mathtt{r}), \qquad \forall \boldsymbol{\theta} \in \Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} \colon \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\}.$$

Theorem

Suppose (ED_0) *and* (ED_2) *,* $(\mathcal{L}_0)$ *,* $(\mathcal{L})$ *, and* $(\mathcal{I})$ *. Let* $\mathtt{b}(\mathtt{r})$ *in* $(\mathcal{L})$ *satisfy*

$$\mathtt{b}(\mathtt{r})\,\mathtt{r} \geq 2z(B, \mathtt{x}) + 2\varrho(\mathtt{r}, \mathtt{x}), \quad \mathtt{r} > \mathtt{r}_0,$$

where $B = D_0^{-1} V_0^2 D_0$ *and*

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0\, z_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\,\omega. \tag{13}$$

Then

$$I\!P\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr) \leq 3e^{-\mathtt{x}}.$$

$(\mathcal{L}G)$ *For each* $\mathtt{r}$ *, there exists* $\mathtt{b}_G(\mathtt{r}) > 0$ *such that* $\mathtt{r}\mathtt{b}_G(\mathtt{r}) \to \infty$ *as* $\mathtt{r} \to \infty$ *and*

$$\frac{-2\mathbb{E} L_G(\boldsymbol{\theta}, \boldsymbol{\theta}_G^*)}{\|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\|^2} \geq \mathtt{b}_G(\mathtt{r}), \qquad \forall \boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r}) = \bigl\{\boldsymbol{\theta} \colon \|D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^*)\| \leq \mathtt{r}\bigr\}.$$

Theorem

Let $\mathtt{b}_G(\mathtt{r})$ *in* $(\mathcal{L}G)$ *satisfy*

$$\mathtt{b}_G(\mathtt{r})\,\mathtt{r} \geq 2z(B_G, \mathtt{x}) + 2\varrho(\mathtt{r}, \mathtt{x}), \quad \mathtt{r} > \mathtt{r}_0,$$

where $B_G = D_G^{-1} V_0^2 D_G$

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\text{def}}{=} 6\nu_0\, z_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\,\omega. \tag{14}$$

Then

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}}_G \notin \Theta_{0,G}(\mathtt{r}_0)\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

Let a vector process $\mathcal{Y}(\boldsymbol{\upsilon})$ fulfill on $\Upsilon_{\circ}(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon}\colon \|\boldsymbol{\upsilon}\| \le \mathtt{r}\}$

$$\sup_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2 \in I\!R^p :\, \|\boldsymbol{\gamma}_1\|=\|\boldsymbol{\gamma}_2\|=1} \log I\!E \exp\Bigl\{\lambda \boldsymbol{\gamma}_1^{\top} \nabla \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{\gamma}_2\Bigr\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

Theorem

Suppose (ED_2) *. It holds on a random set* $\Omega(\mathtt{r},\mathtt{x})$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| \le 6 \nu_0\, z_{\mathbb{H}}(\mathtt{x})\, \mathtt{r},$$

where the function $z_{\mathbb{H}}(\mathtt{x})$ *is given by*

$$z_{\mathbb{H}}(\mathtt{x}) \;=\; \mathbb{H}_1 + \sqrt{2\mathtt{x}} + \mathtt{g}^{-1}(\mathtt{g}^{-2}\mathtt{x} + 1)\mathbb{H}_2,$$

with $\mathbb{H}_2 = 4p$ *and* $\mathbb{H}_1 = 2p^{1/2}$ *.*

Let a vector process $\mathcal{Y}(\boldsymbol{\upsilon})$ fulfill on $\Upsilon_{\circ}(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon}\colon \|B_G^{-1/2}\boldsymbol{\upsilon}\| \le \mathtt{r}\}$

$$\sup_{\boldsymbol{\gamma}_1,\boldsymbol{\gamma}_2 \in I\!\!R^p :\ \|\boldsymbol{\gamma}_1\|=\|\boldsymbol{\gamma}_2\|=1} \log I\!\!E \exp\Bigl\{\lambda \boldsymbol{\gamma}_1^{\top} \nabla \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{\gamma}_2\Bigr\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

Theorem

Suppose (ED_2) *. It holds on a random set* $\Omega(\mathtt{r},\mathtt{x})$

$$\sup_{\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})} \|B_G^{1/2} \mathcal{Y}(\boldsymbol{\upsilon})\| \le 6 \nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \mathtt{r},$$

where the function $z_{\mathbb{H}}(\mathtt{x})$ *is given by*

$$z_{\mathbb{H}}(\mathtt{x}) \ = \ \mathbb{H}_1 + \sqrt{2\mathtt{x}} + \mathtt{g}^{-1}(\mathtt{g}^{-2}\mathtt{x} + 1)\mathbb{H}_2,$$

with

$$\mathbb{H}_1 = \mathbb{H}_1(B_G) = 1 + 2\sqrt{\operatorname{tr}\bigl(B_G \log(B_G)\bigr)}, \quad \mathbb{H}_2 = \mathbb{H}_2(B) \ = \ 1 + \frac{8}{3} \operatorname{tr}\bigl(B_G^{1/2}\bigr).$$

On $\Omega(\mathtt{r}, \mathtt{x})$, for each $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$

$$\bigl\| D_0^{-1} \bigl\{ \nabla I\!E L(\boldsymbol{\theta}) - \nabla I\!E L(\boldsymbol{\theta}^*) \bigr\} + D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr\| \le \delta(\mathtt{r}) \mathtt{r},$$

$$\| D_0^{-1} \bigl\{ \nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*) \bigr\} \| \le 6 \nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \omega \, \mathtt{r}$$

Theorem

Suppose $(\mathcal{L}_0)$ *and* (ED_2) *on* $\Theta_0(\mathtt{r})$ *for a fixed* $\mathtt{r}$ *. Then on* $\Omega(\mathtt{r}, \mathtt{x})$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl\| D_0^{-1} \bigl\{ \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*) \bigr\} + D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \bigr\| \le \diamondsuit(\mathtt{r}, \mathtt{x}),$$

where

$$\diamondsuit(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} \bigl\{ \delta(\mathtt{r}) + 6 \nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \omega \bigr\} \mathtt{r}.$$

The dimension p enters only via the entropy $\mathbb{H}$ in $z_{\mathbb{H}}(\mathtt{x})$.

Local linear approximation of the gradient "penalized"

On $\Omega(\mathtt{r},\mathtt{x})$, for each $\boldsymbol{\theta} \in \Theta_{0,G}(\mathtt{r})$

$$\left\| D_G^{-1}\{\nabla I\!E L_G(\boldsymbol{\theta}) - \nabla I\!E L_G(\boldsymbol{\theta}_G^{*})\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^{*})\right\| \leq \delta_G(\mathtt{r})\mathtt{r},$$

$$\left\| D_G^{-1}\{\nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}_G^{*})\}\right\| \leq 6\nu_0\, z_{\mathbb{H}}(\mathtt{x})\,\omega\,\mathtt{r}$$

Theorem

Suppose $(\mathcal{L}_0 G)$ *and* $(ED_2 G)$ *on* $\Theta_{0,G}(\mathtt{r})$ *for a fixed* $\mathtt{r}$. *Then on* $\Omega(\mathtt{r},\mathtt{x})$

$$\sup_{\boldsymbol{\theta}\in\Theta_{0,G}(\mathtt{r})} \left\| D_G^{-1}\{\nabla L_G(\boldsymbol{\theta}) - \nabla L_G(\boldsymbol{\theta}_G^{*})\} + D_G(\boldsymbol{\theta} - \boldsymbol{\theta}_G^{*})\right\| \leq \diamondsuit_G(\mathtt{r},\mathtt{x}),$$

where

$$\diamondsuit_G(\mathtt{r},\mathtt{x}) \stackrel{\text{def}}{=} \bigl\{\delta_G(\mathtt{r}) + 6\nu_0\, z_{\mathbb{H}}(\mathtt{x})\,\omega\bigr\}\mathtt{r}.$$

The effective dimension $\mathtt{p}_G$ enters only via the entropy $\mathbb{H}$ in $z_{\mathbb{H}}(\mathtt{x})$.

Let $p = p_n \to \infty$. We know

$$\diamondsuit_n(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p_n + \mathtt{x})^2}{n}}, \qquad \Delta_n(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(p_n + \mathtt{x})^3}{n}}, \qquad \|\boldsymbol{\xi}_n\|^2 \le p_n + \mathtt{C}\mathtt{x}.$$

- $p_n/n \to 0$: Consistency:

$$\|\sqrt{\mathbb{F}_{\boldsymbol{\theta}^*}}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*)\| = n^{-1/2}\{\|\boldsymbol{\xi}_n\| \pm \diamondsuit_n(\mathtt{x})\} \le \mathtt{C}\sqrt{\frac{p_n + \mathtt{x}}{n}} \pm \mathtt{C}\,\frac{p_n + \mathtt{x}}{n}$$

- $p_n^2/n \to 0$ – Fisher expansion, root- n normality;

$\sqrt{n\mathbb{F}_{\boldsymbol{\theta}^*}}(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) = \boldsymbol{\xi}_n \pm \diamondsuit_n(\mathtt{x}),$	Expansion of the MLE
$\sqrt{2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)} = \|\boldsymbol{\xi}_n\| \pm 3\diamondsuit_n(\mathtt{x}),$	square-root maximum likelihood
$p_n^{-1/2}L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) = p_n^{-1/2}\|\boldsymbol{\xi}_n\|^2/2 \pm \mathtt{C}\diamondsuit_n(\mathtt{x}),$	likelihood ratio tests, model selection

- $p_n^3/n \to 0$ – Wilks approximation, BvM Theorem.

Let $p = p_n \to \infty$. We know

$$\diamondsuit_G(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(\mathtt{p}_G + \mathtt{x})^2}{n}}, \qquad \Delta_G(\mathtt{x}) \le \mathtt{C}\sqrt{\frac{(\mathtt{p}_G + \mathtt{x})^3}{n}}, \qquad \|\boldsymbol{\xi}_G\|^2 \le \mathtt{p}_G + \mathtt{C}\mathtt{x}.$$

- $\mathtt{p}_G/n \to 0$: Consistency: with $\mathbb{F}_G = \mathbb{F}_{\boldsymbol{\theta}_G^*} + n^{-1}G^2$

$$\|\sqrt{\mathbb{F}_G}(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*)\| = n^{-1/2}\{\|\boldsymbol{\xi}_G\| \pm \diamondsuit_G(\mathtt{x})\} \le \mathtt{C}\sqrt{\frac{\mathtt{p}_G + \mathtt{x}}{n}} \pm \mathtt{C}\,\frac{\mathtt{p}_G + \mathtt{x}}{n}$$

- $\mathtt{p}_G^2/n \to 0$ – Fisher expansion, root- n normality;

$$\sqrt{n\mathbb{F}_G}(\widetilde{\boldsymbol{\theta}}_G - \boldsymbol{\theta}_G^*) = \boldsymbol{\xi}_G \pm \diamondsuit_G(\mathtt{x}),$$ Expansion of the MLE

$$\sqrt{2L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*)} = \|\boldsymbol{\xi}_G\| \pm 3\diamondsuit_G(\mathtt{x}),$$ square-root maximum likelihood

$$\mathtt{p}_G^{-1/2} L_G(\widetilde{\boldsymbol{\theta}}_G, \boldsymbol{\theta}_G^*) = \mathtt{p}_G^{-1/2}\|\boldsymbol{\xi}_G\|^2/2 \pm \mathtt{C}\diamondsuit_G(\mathtt{x}),$$ likelihood ratio tests, model selection

- $\mathtt{p}_G^3/n \to 0$ – Wilks approximation, BvM Theorem.

Outline

1 Bernstein – von Mises Theorem
- BvM Theorem
- Local Gaussian approximation of the posterior
- Tail posterior probability and contraction
- Credible sets

2 Semiparametric estimation
- Motivation
- Linear models
- General semiparametric setup

3 Penalized MLE and effective dimension
- Curse of dimension
- Effective dimension
- Fisher and Wilks expansions
- Concentration and large deviations
- A bound for the norm of a vector stochastic process

4 Confidence estimation using bootstrap
- Likelihood-based confidence set
- Multipier bootstrap
- Conditions

Likelihood-based confidence set

The $1-\alpha$ confidence set for $\boldsymbol{\theta}^*$:

$$\mathcal{E}(\mathfrak{z}_\alpha) \stackrel{\text{def}}{=} \{\boldsymbol{\theta}\colon L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le \mathfrak{z}_\alpha\},$$

$$I\!\!P\left(\boldsymbol{\theta}^* \notin \mathcal{E}(\mathfrak{z}_\alpha)\right) \le \alpha.$$

For the known $L(\boldsymbol{\theta})$ and α the set is determined by the critical value $\mathfrak{z}_\alpha$, the $1-\alpha$ quantile of the excess $L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)$.

For $L(\boldsymbol{\theta}) = -\|\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}\|^2/2$, $\mathcal{E}(\mathfrak{z})$ is an ellipsoid:

$$\mathcal{E}(\mathfrak{z}) = \{\boldsymbol{\theta}\colon \|\Psi^\top(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\|^2 \le 2\mathfrak{z}\}.$$

► Under PA, in the asymptotic setup, $\mathfrak{z}_\alpha$ is close to $1-\alpha$ quantiles of χ^2_p due to the Wilks phenomenon:

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}_n\|^2/2, \qquad \boldsymbol{\xi}_n \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_p), \quad n \to \infty.$$

► But the speed of convergence is slow and under PA-PW the limit distribution is non-pivotal, i.e. depends on $I\!P$.

► The non-asymptotic Wilks result cannot help directly, since the deviation bound for $\|\boldsymbol{\xi}\|^2$ is also non-pivotal and is too rough for a sharp confidence set

$$\Bigl| L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2/2 \Bigr| \le \Delta(\mathrm{x}),$$

$$I\!P\left(\|\boldsymbol{\xi}\|^2 \ge \mathtt{C}(p + 6\mathrm{x})\right) \le 2\mathrm{e}^{-\mathrm{x}}.$$

The idea is to mimic the distribution of $L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*)$ using multiplier bootstrap.

Below $\ell_i(\boldsymbol{\theta})$ is the log-density of Y_i : $\ell_i(\boldsymbol{\theta}) = \log \frac{dP_i(\boldsymbol{\theta})}{d\mu_0}(Y_i)$ and

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}).$$

- Take an i.i.d. sample $u_1, \ldots, u_n$ independent of the data $\boldsymbol{Y}$, $\mathbb{E}(u_i) = \mathrm{Var}(u_i) = 1$ (e.g. $u_i \sim \exp(1)$ or $\mathcal{N}(1,1)$).
- Bootstrap the likelihood function:

$$L^{\circ}(\boldsymbol{\theta}) = L^{\circ}(\boldsymbol{\theta}, \boldsymbol{u}) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})\, u_i$$

$^{\circ}$ denotes the conditional probability with the fixed sample $\boldsymbol{Y}$.

"$\boldsymbol{Y}$ world"	"bootstrap world"
MLE	
$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$	$\widetilde{\boldsymbol{\theta}}^{\circ} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta})$
target	
$\boldsymbol{\theta}^{*} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta})$	$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname{argmax}_{\boldsymbol{\theta}} I\!E^{\circ} L^{\circ}(\boldsymbol{\theta})$
likelihood ratio	
$L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^{*})$	$L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}) - L^{\circ}(\widetilde{\boldsymbol{\theta}})$

- The bootstrap side is fully computable!
- The true point in bootstrap world is exactly qMLE $\widetilde{\boldsymbol{\theta}}$.
- The "bootstrap world" is built inside of the parametric model, which may be wrong.

Questions to be addressed:

- Bootstrap consistency in non-asymptotic form
- Error of coverage probability
- Size of the bootstrap-based confidence set

Key ingredients:

- Fisher and Wilks expansions in real and bootstrap worlds;
- Closeness of distributions of the of approximating terms $\|\boldsymbol{\xi}\|^2$ and $\|\boldsymbol{\xi}^{\circ}\|^2$;
- Closeness of the local metrics on the parameter space:

$$D_0^2 \approx \mathfrak{D}_0^2 \quad \Leftrightarrow \quad \nabla_{\boldsymbol{\theta}}^2 \mathbb{E} L(\boldsymbol{\theta}^*) \approx \nabla_{\boldsymbol{\theta}}^2 \mathbb{E}^{\circ} L^{\circ}(\widetilde{\boldsymbol{\theta}});$$

- Use of the truncated moment-generating function to get a sharp bound for

$$\mathcal{L}(\|\boldsymbol{\xi}\|^2) \approx \mathcal{L}(\|\boldsymbol{\xi}^{\circ}\|^2 \,|\, \boldsymbol{Y}).$$

Theorem

It holds with $I\!P^{\circ}-$ *probability* $\geq 1 - C\mathrm{e}^{-\mathtt{x}}$.

$$\Bigl| L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}, \widetilde{\boldsymbol{\theta}}) - \|\boldsymbol{\xi}^{\circ}\|^2/2 \Bigr| \leq \varDelta^{\circ}(\mathtt{x}),$$

where the following terms are $I\!P-$ *random*

$$\boldsymbol{\xi}^{\circ} \stackrel{\mathrm{def}}{=} \mathfrak{D}_0^{-1} \nabla_{\boldsymbol{\theta}} L^{\circ}(\widetilde{\boldsymbol{\theta}}), \qquad \mathfrak{D}_0^2 \stackrel{\mathrm{def}}{=} -\nabla_{\boldsymbol{\theta}}^2 I\!E^{\circ} L^{\circ}(\widetilde{\boldsymbol{\theta}}).$$

Two Wilks results lead to the following scheme:

$$L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \approx \|\boldsymbol{\xi}\|^2/2$$

$$\wr\wr$$

$$L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}, \widetilde{\boldsymbol{\theta}}) \approx \|\boldsymbol{\xi}^{\circ}\|^2/2.$$

The Wilks theorems results are valid on two different probability spaces. The approximation $\approx$ connects two "worlds" in distribution:

$$\mathcal{L}(\|\boldsymbol{\xi}\|^2) \approx \mathcal{L}(\|\boldsymbol{\xi}^{\circ}\|^2 \,|\, \boldsymbol{Y}).$$

Leading to

$$\mathcal{L}\{L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*)\} \approx \mathcal{L}\{L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}, \widetilde{\boldsymbol{\theta}}) \,|\, \boldsymbol{Y}\}.$$

Theorem

Let the conditions (ED_2) , (ED_3) *and* $(\mathcal{L}_0)$ *be fulfilled, then it holds with probability* $\geq 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\sup_{\substack{\boldsymbol{\gamma}_{1,2} \in I\!R^p, \\ \|\boldsymbol{\gamma}_{1,2}\|=1}} \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} \left| \boldsymbol{\gamma}_1^\top D_0^{-1} \mathfrak{D}^2(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 - 1 \right| \leq \mathtt{C} \sqrt{(p + \mathtt{x})^3 / n},$$

where

$$\mathfrak{D}^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta}), \qquad D_0^2 \stackrel{\text{def}}{=} -\sum_{i=1}^{n} I\!E \nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta}^*).$$

This result implies that on the set $\Omega(\mathtt{x})$ of a dominating probability $1 - \mathtt{C}\,\mathrm{e}^{\mathtt{x}}$

$$\| D_0^{-1} \mathfrak{D}_0^2 D_0^{-1} - \boldsymbol{I}_p \|_\infty \leq \mathtt{C} \sqrt{(p + \mathtt{x})^3 / n}.$$

Lemma

It holds with $\mathbb{P}^{\circ}$*– probability* $\geq 1 - 2\mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta_0^{\circ}(\mathtt{r}_0)} \|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}_1) - \boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}_2)\| \leq \mathtt{C}(p + \mathtt{x})/\sqrt{n}.$$

Moreover

$$\left| \|\boldsymbol{\xi}^{\circ}(\widetilde{\boldsymbol{\theta}})\|^2 - \|\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}^*)\|^2 \right| \leq \mathtt{C}\sqrt{(p + \mathtt{x})^3/n},$$

where

$$\boldsymbol{\xi}^{\circ}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathfrak{D}_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} L^{\circ}(\boldsymbol{\theta}) - \nabla_{\boldsymbol{\theta}} \mathbb{E}^{\circ} L^{\circ}(\boldsymbol{\theta}) \right\},$$

$$\Theta_0^{\circ}(\mathtt{r}_0) \stackrel{\text{def}}{=} \left\{ \boldsymbol{\theta} \colon \|\mathfrak{D}_0(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})\| \leq \mathtt{r}_0 \right\}.$$

Approximating terms

Remind the definition:

Normalized score functions:

$$\boldsymbol{\xi} = D_0^{-1} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) = D_0^{-1} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*),$$

$$\boldsymbol{\xi}^{\circledcirc} = \mathfrak{D}_0^{-1} \nabla_{\boldsymbol{\theta}} L^{\circledcirc}(\widetilde{\boldsymbol{\theta}}) = \mathfrak{D}_0^{-1} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\widetilde{\boldsymbol{\theta}}) (\boldsymbol{u_i} - \mathbf{1}).$$

Fisher Information matrices

$$D_0^2 = - \sum_{i=1}^{n} \mathbb{E} \nabla_{\boldsymbol{\theta}}^2 \ell_i(\boldsymbol{\theta}^*) \quad \text{deterministic},$$

$$\mathfrak{D}_0^2 = - \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}^2 \ell_i(\widetilde{\boldsymbol{\theta}}) \quad \mathbb{P} - \text{random}.$$

Approximating terms

Due to the previous results one can make the following substitution: on a set of probability $\geq 1 - \mathtt{C}\mathtt{e}^{-\mathtt{x}}$:

$$\mathfrak{D}_0^2 \approx D_0^2, \qquad \|\boldsymbol{\xi}^{\circ}(\widetilde{\theta})\|^2 \approx \|\boldsymbol{\xi}^{\circ}(\theta^*)\|^2, \qquad \boldsymbol{\xi}^{\circ}(\widetilde{\theta}) \approx \boldsymbol{\xi}^{\circ}(\theta^*).$$

$$\boldsymbol{\xi}^{\circ}(\widetilde{\theta}) \approx \boldsymbol{\xi}^{\circ}(\theta^*) \;\approx\; D_0^{-1} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \boldsymbol{(u_i - 1)},$$

$$\boldsymbol{\xi} \stackrel{\text{def}}{=} D_0^{-1} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*).$$

Multiplier CLT [van der Vaart and Wellner, 1996]

In the i.i.d. case with the true parametric model it holds

$$V_0^{-1} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)(u_i - 1) \xrightarrow{\mathcal{L}^{\circ}} \mathcal{N}(0, \boldsymbol{I}_p),$$

for almost every i.i.d. sequence $u_1, u_2 \ldots$ s.t. $\mathbb{E}^{\circ} u_i = 1, \mathrm{Var}^{\circ} u_i = 1$, with

$$\begin{aligned} V_0^2 &\stackrel{\mathrm{def}}{=} \mathrm{Var}\{\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*)\} \\ &= D_0^2 \quad \text{for the true parametric model.} \end{aligned}$$

Therefore, in the i.i.d. parametric case the approximating vectors $\boldsymbol{\xi}$ and $\boldsymbol{\xi}^{\circ} \approx \boldsymbol{\xi}^{\circ}(\theta^*)$ have the same limit distributions.

Introduce for $\varepsilon \sim \mathcal{N}(0, \boldsymbol{I}_p)$, fixed $\mathit{\Gamma}_0 = \mathtt{C}\sqrt{p}$ and arbitrary $\boldsymbol{\gamma} \in I\!\!R^p$, $\|\boldsymbol{\gamma}\| = 1$:

$$h(\mu, t) \stackrel{\text{def}}{=} \exp(\mu t/2) I\!\!P\left(\|\boldsymbol{\varepsilon} + \sqrt{\mu t}\boldsymbol{\gamma}\| \le \mu^{-1/2}\mathit{\Gamma}_0\right).$$

Theorem

It holds with probability $\geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$

$$\sup_{\mu \in (0,1)} \left| \frac{I\!\!E^{\circ} h(\mu, \|\boldsymbol{\xi}^{\circ}\|^2)}{I\!\!E h(\mu, \|\boldsymbol{\xi}\|^2)} - 1 \right| \le \mathtt{C}\sqrt{\frac{(p + \mathtt{x})^3}{n}}.$$

Idea of the proof:

Get to the linear exponent w.r.t. $\boldsymbol{\xi}$ by

$$\exp\bigl(\mu\|\boldsymbol{\xi}\|^2/2\bigr)\,I\!\!P\Bigl(\|\boldsymbol{\varepsilon}+\mu^{1/2}\boldsymbol{\xi}\|\le\mu^{-1/2}\mathit{\Gamma}_0\,\big|\,\boldsymbol{\xi}\Bigr)$$

$$=\frac{1}{(2\pi\mu)^{p/2}}\int_{\|\boldsymbol{\gamma}\|\le\mathit{\Gamma}_0}\exp\Bigl(\boldsymbol{\gamma}^\top\boldsymbol{\xi}-\frac{1}{2\mu}\|\boldsymbol{\gamma}\|^2\Bigr)\,d\boldsymbol{\gamma}.$$

Use the Taylor expansion of $\log I\!\!E\exp\bigl(\lambda\boldsymbol{\gamma}^\top\boldsymbol{\xi}\bigr)$ w.r.t. $|\lambda|\le\mathit{\Gamma}_0=\mathtt{C}\sqrt{p}$.

A cumulative bound

Let $\mathfrak{z}_{\alpha}^{\circ}$ denote the upper α -quantile of $L^{\circ}(\widetilde{\boldsymbol{\theta}}^{\circ}, \widetilde{\boldsymbol{\theta}})$.

Theorem

It holds with probability $\geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{y}}$

$$\mathbb{P}\left(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) > \mathfrak{z}_{\alpha}^{\circ} + \Delta_{cum}\right) - \alpha \leq \alpha\delta_{F},$$

$$\mathbb{P}\left(L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^{*}) > \mathfrak{z}_{\alpha}^{\circ} - \Delta_{cum}\right) - \alpha \geq -\alpha\delta_{F},$$

where

$$\Delta_{cum}, \, \delta_{F} \lesssim \sqrt{\frac{(p + \mathtt{y})^{3}}{n}}.$$

Squared modelling bias and a size of the confidence set

Compare the approximating terms $I\!E\|\boldsymbol{\xi}\|^2$ and $I\!E^{\circ}\|\boldsymbol{\xi}^{\circ}\|^2$:

$$I\!E\|\boldsymbol{\xi}\|^2 = \operatorname{tr}\bigl(D_0^{-1} V_0^2 D_0^{-1}\bigr), \qquad I\!E^{\circ}\|\boldsymbol{\xi}^{\circ}\|^2 = \operatorname{tr}\bigl(\mathfrak{D}_0^{-1} \mathcal{V}_0^2 \mathfrak{D}_0^{-1}\bigr).$$

$$\begin{aligned} V_0^2 &\stackrel{\text{def}}{=} \operatorname{Var} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}^*) \\ &= \sum_{i=1}^n I\!E\Bigl[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top\Bigr] - \sum_{i=1}^n I\!E\bigl[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)\bigr] I\!E\bigl[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)\bigr]^\top, \\ \mathcal{V}_0^2 &\stackrel{\text{def}}{=} \operatorname{Var}^{\circ} \nabla_{\boldsymbol{\theta}} L^{\circ}(\widetilde{\boldsymbol{\theta}}) \\ &= \sum_{i=1}^n \nabla_{\boldsymbol{\theta}} \ell_i(\widetilde{\boldsymbol{\theta}}) \nabla_{\boldsymbol{\theta}} \ell_i(\widetilde{\boldsymbol{\theta}})^\top . \end{aligned}$$

The relation of the blue matrices in spectral norm is $\le \mathtt{C}\sqrt{(p+\mathtt{x})^3/n}$. The magenta matrix adds the modelling bias, bounded by condition (SmB) .

(ED_3) *It holds for all* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *,* $\mathtt{r} \le \mathtt{r}_0$ *and for* $j = 1, 2, 3$ and $|\lambda| \le \mathtt{g}$

$$\sup_{\substack{\boldsymbol{\gamma}_j \in I\!R^p, \\ \|\boldsymbol{\gamma}_j\| \le 1}} \log I\!E \exp\left\{ \frac{\lambda}{\omega_1} \boldsymbol{\gamma}_3^\top \nabla_{\boldsymbol{\theta}} \left[\boldsymbol{\gamma}_1^\top D_0^{-1} \nabla_{\boldsymbol{\theta}}^2 \zeta(\boldsymbol{\theta}) D_0^{-1} \boldsymbol{\gamma}_2 \right] \right\} \le \frac{\nu_0^2 \lambda^2}{2}.$$

(SmB) *There exists a constant* $\delta_\xi^2 \lesssim \sqrt{p/n^3}$ *such that it holds for all* $i = 1, \ldots, n$

$$\left\| D_0^{-1} I\!E \nabla_{\boldsymbol{\theta}} \log \frac{dP_i(\boldsymbol{\theta}^*)}{d\mu_0}(Y_i) \right\| \le \delta_\xi$$

(SD_0) *There exists a constant $\delta_v \geq 0$ such that it holds for all $i = 1, \ldots, n$ with dominating probability*

$$\left\| H_0^{-1} \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top - I\!\!E \left[\nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right] \right\} H_0^{-1} \right\| \leq \delta_v,$$

where

$$H_0^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} I\!\!E \left\{ \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*) \nabla_{\boldsymbol{\theta}} \ell_i(\boldsymbol{\theta}^*)^\top \right\}.$$

(Condition for the non-commutative Bernstein inequality by [Koltchinskii et al., 2011])

Bickel, P. J. and Kleijn, B. J. K. (2012).
The semiparametric Bernstein-von Mises theorem.
Ann. Statist., 40(1):206–237.

Bontemps, D. (2011).
Bernstein-von Mises theorems for Gaussian regression with increasing number of regressors.
Ann. Statist., 39(5):2557–2584.

Boucheron, S. and Gassiat, E. (2009).
A Bernstein-von Mises theorem for discrete probability distributions.
Electron. J. Stat., 3:114–148.

Bunke, O. and Milhaud, X. (1998).
Asymptotic behavior of Bayes estimates under possibly incorrect models.
Ann. Statist., 26(2):617–644.

Castillo, I. (2012).
A semiparametric Bernstein - von Mises theorem for Gaussian process priors.
Probability Theory and Related Fields, 152:53–99.
10.1007/s00440-010-0316-5.

Castillo, I. and Nickl, R. (2013).
Nonparametric Bernstein–von Mises theorems in Gaussian white noise.
Ann. Statist., 41(4):1999–2028.

Castillo, I. and Rousseau, J. (2013).
A general bernstein–von mises theorem in semiparametric models.

Cox, D. D. (1993).
An analysis of Bayesian inference for nonparametric regression.
Ann. Stat., 21(2):903–923.

Freedman, D. (1999).
On the Bernstein-von Mises theorem with infinite-dimensional parameters.
Ann. Stat., 27(4):1119–1140.

Ghosal, S. (1999).
Asymptotic normality of posterior distributions in high-dimensional linear models.
Bernoulli, 5(2):315–331.

Ghosal, S. (2000).
Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity.

J. Multivariate Anal., 74(1):49–68.

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000).
Convergence rates of posterior distributions.
Ann. Statist., 28(2):500–531.

Ghosal, S. and van der Vaart, A. (2007).
Convergence rates of posterior distributions for noniid observations.
Ann. Statist., 35:192.

Kim, Y. (2006).
The Bernstein-von Mises theorem for the proportional hazard model.
Ann. Statist., 34(4):1678–1700.

Kleijn, B. J. K. and van der Vaart, A. W. (2006).
Misspecification in infinite-dimensional Bayesian statistics.
Ann. Statist., 34(2):837–877.

Kleijn, B. J. K. and van der Vaart, A. W. (2012).
The Bernstein-von-Mises theorem under misspecification.
Electronic J. Statist., 6:354–381.

Koltchinskii, V., Lounici, K., and Tsybakov, A. B. (2011).

Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion.
Ann. Stat., 39(5):2302–2329.

Rivoirard, V. and Rousseau, J. (2012).
Bernstein–von mises theorem for linear functionals of the density.
Ann. Stat., 40(3):1489–1523.

Shen, X. (2002).
Asymptotic normality of semiparametric and nonparametric posterior distributions.
J. Amer. Statist. Assoc., 97(457):222–235.

van der Vaart, A. and Wellner, J. A. (1996).
Weak convergence and empirical processes. With applications to statistics.
Springer Series in Statistics. New York, Springer.