Institute for Information Transmission Problems

PreMoLab
Moscow Institute of
Physics and Technology

DATADVANCE
AN EADS COMPANY

Semiparametric Bernstein - von Mises theorem: non-asymptotic approach.

Maxim Panov, Vladimir Spokoiny

Outline

Data $Y = (Y_1, \ldots, Y_n) \sim \mathbb{P}$.

PA: $\mathbb{P} \in (\mathbb{P}_\upsilon, \upsilon \in \Upsilon \subseteq \mathbb{R}^p)$.

$$L(\upsilon) = L(Y, \upsilon) \stackrel{\text{def}}{=} \log \frac{d\mathbb{P}_\upsilon}{d\mu_0}(Y) = \log p(Y \,\big|\, \upsilon).$$

Bayes approach: υ is a random element on Υ from a prior π:

$$Y \mid \upsilon \ \sim \ p(y \,\big|\, \upsilon) = \exp L(\upsilon),$$

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

Target: the *posterior* distribution of υ given Y:

$$\upsilon \mid Y \ \propto \ \exp\bigl\{L(\upsilon)\bigr\}\, \pi(d\upsilon) = \exp\bigl\{L(\upsilon)\bigr\}\, \pi(\upsilon) d\upsilon.$$

Maximum Likelihood Estimation (MLE)

$$\tilde{\upsilon} = \operatorname*{argmax}_{\upsilon \in \Upsilon} L(\upsilon).$$

$\tilde{\upsilon}$ concentrates around the “true” parameter

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

Fisher Theorem

$$\sqrt{D^2(\upsilon^*)}(\tilde{\upsilon} - \upsilon^*) \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_p)$$

where $D^2(\upsilon^*)$ is the Fisher information matrix at υ^*, $p = \dim(\Upsilon)$.

Bernstein-von Mises Theorem

$$\sqrt{D^2(\upsilon^*)}(\upsilon - \tilde{\upsilon}) \mid Y \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_p),$$

Aim: concentration of the posterior, near Gaussianity.

The use of BvM

- Given data Y, the posterior measure is given

$$\upsilon \mid Y \propto \exp\{L(\upsilon)\}\, \pi(d\upsilon).$$

- Using Bayes computation, one can draw a sample from the posterior and estimate its mean and variance.

- BvM Theorem states near normality of the posterior.

- one can build elliptic credible sets $\mathcal{A}_\alpha$ on which the posterior concentrates with a prescribed probability $1-\alpha$.

The use of BvM

Theorem (Bernstein - von Mises)

For any Borel set $\mathcal{A}$, *with* $\mathcal{D}^2 = -\nabla^2 \mathbb{E} L(\upsilon^*)$

$$\mathbb{P}\bigl(\mathcal{D}(\upsilon - \tilde{\upsilon}) \in \mathcal{A} \,|\, Y\bigr) \;\approx\; \Phi(\mathcal{A}).$$

Corollary (Moments of the posterior)

Define $\overline{\upsilon} \stackrel{\mathrm{def}}{=} \mathbb{E}\bigl(\upsilon \,\big|\, Y\bigr)$ *and* $\mathfrak{S}^2 \stackrel{\mathrm{def}}{=} \mathrm{Cov}\bigl(\upsilon \,\big|\, Y\bigr)$. *It holds*

$$\mathcal{D}^2 \mathfrak{S}^2 \approx \boldsymbol{I}_p, \qquad \mathcal{D}\bigl(\overline{\upsilon} - \tilde{\upsilon}\bigr) \approx 0, \qquad \mathfrak{S}^{-1}\bigl(\upsilon - \overline{\upsilon}\bigr) \approx \mathcal{N}(0, \boldsymbol{I}_p).$$

Corollary (Credible sets)

With $\mathcal{A}_\alpha = \bigl\{\upsilon : \|\mathfrak{S}^{-1}(\upsilon - \overline{\upsilon})\| \le z_\alpha\bigr\}$, *it holds*

$$\mathbb{P}\bigl(\upsilon^* \notin \mathcal{A}_\alpha\bigr) \approx \mathbb{P}(\|\gamma\| > z_\alpha),$$

where γ *is a standard Gaussian vector in* $\mathbb{R}^p$ *and* z_α^2 *is a quantile of chi-scuared distribution with* p *degrees of freedom.*

Finite sample Bayesian theory of Spokoiny (Baldin, P., ...)

Goal: An extension of the BvM result for following situations:

- finite samples
- large parameter dimension
- model (likelihood) misspecification
- Gaussian (regularization) priors
- BvM for hyperpriors and Bayesian model selection

Outline

BvM Theorem: General approach

$$Y \mid \upsilon \sim \exp L(y|\upsilon),$$
$$\upsilon \sim \pi(\cdot).$$

Define

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

Aim: to bound the probability of $\{\upsilon \in \mathcal{A}\}$ given Y for $\mathcal{A} \subset \Upsilon$:

$$I\!P\bigl(\upsilon \in \mathcal{A} \mid Y\bigr) = \frac{\int \exp\bigl\{L(\upsilon)\bigr\}\, \mathbf{I}\bigl\{\upsilon \in \mathcal{A}\bigr\} d\pi(\upsilon)}{\int \exp\bigl\{L(\upsilon)\bigr\} d\pi(\upsilon)}$$

Step 1: Concentration of the posterior in a local neighborhood Υ_0 of υ^*:

$$\int \exp\bigl\{L(\upsilon)\bigr\}\, \mathbf{I}\bigl\{\upsilon - \upsilon^* \in \Upsilon \setminus \Upsilon_0\bigr\} d\pi(\upsilon) \ll \int \exp\bigl\{L(\upsilon)\bigr\}\, \mathbf{I}\bigl\{\upsilon - \upsilon^* \in \Upsilon_0\bigr\} d\pi(\upsilon)$$

Local quadratic approximation

Step 2: Local quadratic approximation

$$\begin{aligned}\mathbb{L}(\upsilon,\upsilon^*) &= (\upsilon-\upsilon^*)^\top \nabla L(\upsilon^*) - \|\mathcal{D}_0(\upsilon-\upsilon^*)\|^2/2, \\ &= (\mathcal{D}_0(\upsilon-\upsilon^*))^\top \xi - \|\mathcal{D}_0(\upsilon-\upsilon^*)\|^2/2\end{aligned}$$

where $\mathcal{D}_0 \stackrel{\text{def}}{=} -\nabla^2 I\!E L(\upsilon^*)$ and $\xi \stackrel{\text{def}}{=} \mathcal{D}_0^{-1}\nabla L(\upsilon^*)$.

Define excess:

$$L(\upsilon,\upsilon^*) = L(\upsilon) - L(\upsilon^*).$$

Theorem

Let for some $\mathtt{r}_0 > 0$ fixed. Then under some generic conditions on local likelihood behavior with dominating probability $\geq 1-\mathrm{e}^{-\mathtt{x}}$

$$|L(\upsilon,\upsilon^*) - \mathbb{L}(\upsilon,\upsilon^*)| \leq \Delta(\mathtt{r}_0,\mathtt{x}), \qquad \upsilon \in \Upsilon_0(\mathtt{r}_0),$$

where $\Upsilon_0(\mathtt{r}_0) = \{\upsilon : \|\mathcal{D}_0(\upsilon-\upsilon^*)\| \leq \mathtt{r}_0\}$ and the random variable $\Delta(\mathtt{r}_0,\mathtt{x})$ is typically small.

The proper choice of $\mathtt{r}_0$ ensures $\Delta(\mathtt{r}_0,\mathtt{x}) \sim \sqrt{\frac{(p+\mathtt{x})^3}{n}}$.

Conditions

The conditions include:

- A growth condition: for some $\mathtt{b} > 0$ and for all υ

$$\mathbb{E}\mathcal{L}(\upsilon) - \mathbb{E}\mathcal{L}(\upsilon^*) \le -\mathtt{b}\|\mathcal{D}_0(\upsilon - \upsilon^*)\|^2.$$

- Exponential moment conditions of the kind that $\zeta(\upsilon) = \mathcal{L}(\upsilon) - \mathbb{E}\mathcal{L}(\upsilon)$ satisfies with some $\nu_0, \mathrm{g} > 0$ and small $\omega > 0$ for any $0 < |\mu| \le \mathrm{g}$ and any $\gamma \in \mathbb{R}^p$ with $\|\gamma\| = 1$

$$\log \mathbb{E} \exp\left\{ \mu \frac{\gamma^\top [\nabla\zeta(\upsilon) - \nabla\zeta(\upsilon^*)]}{\omega\|\mathcal{D}_0(\upsilon - \upsilon^*)\|} \right\} \le \nu_0 \mu^2/2.$$

- A smoothness condition: for $\upsilon \in \Upsilon_0(\mathtt{r}_0)$

$$\|\nabla^2\mathbb{E}\mathcal{L}(\upsilon) - \nabla^2\mathbb{E}\mathcal{L}(\upsilon^*)\| \le \delta(\mathtt{r}_0)\mathtt{r}_0^2.$$

Then, error of quadratic approximation becomes

$$\Delta(\mathtt{r}_0, \mathtt{x}) \stackrel{\text{def}}{=} \{\delta(\mathtt{r}_0) + 6\nu_0 \, p^{1/2}\, \omega\} \, \mathtt{r}_0^2.$$

Local Gaussian approximation for the posterior

Consider non-informative prior $\pi(\upsilon) \equiv 1, \upsilon \in \Upsilon$.

Aim: to bound the probability of $\{\mathcal{D}_0(\upsilon - \upsilon^\circ) \in \mathcal{A}\}$ given Y:

$$
I\!\!P\big(\mathcal{D}_0(\upsilon - \upsilon^\circ) \in \mathcal{A} \mid Y\big) = \frac{\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{L(\upsilon)\} d\upsilon}{\int \exp\{L(\upsilon)\} d\upsilon}
$$

$$
= \frac{\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{L(\upsilon) - L(\upsilon^*)\} d\upsilon}{\int \exp\{\mathbb{L}(\upsilon) - L(\upsilon^*)\} d\upsilon} = \frac{\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{L(\upsilon,\upsilon^*)\} d\upsilon}{\int \exp\{L(\upsilon,\upsilon^*)\} d\upsilon}
$$

where $\upsilon^\circ \stackrel{\text{def}}{=} \upsilon^* + \mathcal{D}_0^{-2} \nabla L(\upsilon^*)$.

By the majorization bound

$$
\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{L(\upsilon,\upsilon^*)\} d\upsilon \le \mathrm{e}^{\Delta(\mathtt{r}_0,\mathtt{x})} \int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{\mathbb{L}(\upsilon,\upsilon^*)\} d\upsilon
$$

Similarly, it holds

$$
\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{L(\upsilon,\upsilon^*)\} d\upsilon \ge \mathrm{e}^{-\Delta(\mathtt{r}_0,\mathtt{x})} \int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}} \exp\{\mathbb{L}(\upsilon,\upsilon^*)\} d\upsilon.
$$

Normalization factors

Define

$$m(\xi) \stackrel{\mathrm{def}}{=} -\|\xi\|^2/2 + \log(\det \mathcal{D}_0) - p\log(\sqrt{2\pi}),$$

then

$$m(\xi) + \mathbb{L}(\upsilon, \upsilon^*) \;=\; -\frac{\|\mathcal{D}_0(\upsilon - \upsilon^*) - \xi\|^2}{2} + \log(\det \mathcal{D}_0) - \frac{p}{2}\log(2\pi)$$

is (conditionally on Y) the log-density of the normal law.

Local Gaussian approximation of the posterior

Change of variables $u = \mathcal{D}_0(\upsilon - \upsilon^\circ)$ implies

$$
\begin{aligned}
&\int \exp\{L(\upsilon, \upsilon^*) + m(\xi)\}\, \mathbf{I}\{\mathcal{D}_0(\upsilon - \upsilon^\circ) \in \mathcal{A}\}\, d\upsilon \\
&\qquad \le \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} \int \exp\{\mathbb{L}(\upsilon, \upsilon^*) + m(\xi)\}\, \mathbf{I}\{\mathcal{D}_0(\upsilon - \upsilon^\circ) \in \mathcal{A}\}\, d\upsilon \\
&\qquad = \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} (2\pi)^{-p/2} \int \exp\bigl(-\|u\|^2/2\bigr)\, \mathbf{I}\bigl(u \in \mathcal{A}\bigr)\, du \\
&\qquad = \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} \mathbb{P}(\gamma \in \mathcal{A}),
\end{aligned}
$$

where γ is a standard normal variable in $\mathbb{R}^p$, and hence,

$$
\int \exp\{L(\upsilon, \upsilon^*)\}\, \mathbf{I}\{\mathcal{D}_0(\upsilon - \upsilon^\circ) \in \mathcal{A}\}\, d\upsilon \le \exp\{\Delta(\mathtt{r}_0, \mathtt{x}) - m(\xi)\}\, \mathbb{P}(\gamma \in \mathcal{A}).
$$

Local Gaussian approximation of the posterior

We further need to bound denominator:

$$\begin{aligned}
&\int \exp\bigl\{L(\upsilon,\upsilon^*)+m(\xi)\bigr\}\,d\upsilon \\
&\quad\geq \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{L(\upsilon,\upsilon^*)+m(\xi)\bigr\}\,d\upsilon \\
&\quad\geq \mathrm{e}^{-\Delta(\mathtt{r}_0,\mathtt{x})} \int_{\Upsilon_0(\mathtt{r}_0)} \exp\bigl\{\mathbb{L}(\upsilon,\upsilon^*)+m(\xi)\bigr\}\,d\upsilon \\
&\quad= \mathrm{e}^{-\Delta(\mathtt{r}_0,\mathtt{x})} I\!P(\|\gamma\|>\mathtt{r}_0),
\end{aligned}$$

where γ is a standard normal random variable in $\mathbb{R}^p$.

Define $\nu(\mathtt{r}_0)=\log I\!P(\|\gamma\|>\mathtt{r}_0)$, and we finally obtain

$$\begin{aligned}
I\!P\bigl(\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}\mid Y\bigr) &= \frac{\int_{\mathcal{D}_0(\upsilon-\upsilon^\circ)\in\mathcal{A}}\exp\bigl\{L(\upsilon)\bigr\}d\upsilon}{\int\exp\bigl\{L(\upsilon)\bigr\}d\upsilon} \\
&\geq \exp\{2\Delta(\mathtt{r}_0,\mathtt{x})-\nu(\mathtt{r}_0)\}\, I\!P(\gamma\in\mathcal{A}).
\end{aligned}$$

Back to step 1: tails of the posterior

Consider

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \exp\{L(\upsilon, \upsilon^*)\} d\upsilon}{\int_{\Upsilon} \exp\{L(\upsilon, \upsilon^*)\} d\upsilon} .$$

Local majorization:

$$L(\upsilon, \upsilon^*) \geq \mathbb{L}(\upsilon, \upsilon^*) - \Delta(\mathtt{r}_0, \mathtt{x}), \qquad \upsilon \in \Upsilon_0(\mathtt{r}_0).$$

Upper function: with probability at least $1 - 2\mathrm{e}^{-\mathtt{x}}$

$$L(\upsilon, \upsilon^*) + \mathfrak{u}(\upsilon) \leq 0, \qquad \upsilon \in \Upsilon \setminus \Upsilon_0(\mathtt{r}_0).$$

Implies

$$\rho(\mathtt{r}_0) \leq \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} \frac{\int_{\Upsilon \setminus \Upsilon_0(\mathtt{r}_0)} \mathrm{e}^{-\mathfrak{u}(\upsilon)} d\upsilon}{\int_{\Upsilon_0(\mathtt{r}_0)} \exp\{\mathbb{L}(\upsilon, \upsilon^*)\} d\upsilon} .$$

Under some conditions $\mathfrak{u}(\upsilon)$ can be chosen as a quadratic function and ensures nice concentration properties.

BvM for non-informative prior

Theorem

Define $\overline{\upsilon} \stackrel{\text{def}}{=} \mathbb{E}\bigl(\upsilon \,\big|\, Y\bigr)$, $\mathfrak{S}^2 \stackrel{\text{def}}{=} \operatorname{Cov}\bigl(\upsilon \,\big|\, Y\bigr)$ *and* $\upsilon^{\circ} = \upsilon^{*} + \mathcal{D}_0^{-1}\xi$,
where $\xi = \mathcal{D}_0^{-1}\nabla L(\upsilon^{*})$.
Then, it holds with dominating probability ($\geq 1 - 4\mathrm{e}^{-\mathtt{x}}$)*:*

$$\bigl\|\mathcal{D}_0(\overline{\upsilon} - \upsilon^{\circ})\bigr\|^2 \leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 4\mathrm{e}^{-\mathtt{x}}$$

$$\bigl\|\boldsymbol{I}_p - \mathcal{D}_0\mathfrak{S}^2\mathcal{D}_0\bigr\|_{\infty} \leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 4\mathrm{e}^{-\mathtt{x}}.$$

Moreover, for any measurable set $\mathcal{A} \subset \mathbb{R}^p$

$$\begin{aligned} &\exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 3\mathrm{e}^{-\mathtt{x}}\bigr)\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr) - \mathrm{e}^{-\mathtt{x}} \\ &\leq \mathbb{P}\bigl(\mathcal{D}_0(\upsilon - \upsilon^{\circ}) \in \mathcal{A} \,\big|\, Y\bigr) \\ &\qquad \leq \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}}\bigr)\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr). \end{aligned}$$

where $\gamma \in \mathbb{R}^p$ *is a standard Gaussian.*

Corollary

Define $\overline{\upsilon} \stackrel{\text{def}}{=} I\!\!E(\upsilon \,|\, Y)$, $\mathfrak{S}^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\upsilon \,|\, Y)$.
Then, it holds with dominating probability ($\geq 1 - 4\mathrm{e}^{-\mathtt{x}}$) *for any measurable set* $\mathcal{A} \subset I\!\!R^p$:

$$\exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 3\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{I\!\!P\bigl(\gamma \in \mathcal{A}\bigr) - \tau\Bigr\} - \mathrm{e}^{-\mathtt{x}}$$
$$\leq I\!\!P\bigl(\mathfrak{S}^{-1}(\upsilon - \overline{\upsilon}) \in \mathcal{A} \,\big|\, Y\bigr)$$
$$\leq \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{I\!\!P\bigl(\gamma \in \mathcal{A}\bigr) + \tau\Bigr\}.$$

where $\tau = \frac{1}{2}\sqrt{(1 + \Delta(\mathtt{r}_0, \mathtt{x}))^2 \Delta(\mathtt{r}_0, \mathtt{x})^2 + \Delta(\mathtt{r}_0, \mathtt{x})^2 p}$ *and* $\gamma \in \mathbb{R}^p$ *is a standard Gaussian.*

This shows that required condition for validity of results changes from "$\Delta(\mathtt{r}_0, \mathtt{x})$ is small" to "$\Delta(\mathtt{r}_0, \mathtt{x}) p^{1/2}$ is small".

Outline

Profile MLE

Semiparametric problem:

$$\upsilon = (\theta, \eta), \qquad \varUpsilon = \varTheta \times H.$$

Goal of estimation:

$$\theta = \varPi_0 \upsilon \qquad \text{for projection} \qquad \varPi_0 : \varUpsilon \to \varTheta, \ \ dim(\varTheta) = q.$$

Profile maximum likelihood (pMLE):

$$\widetilde{\theta} = \varPi_0 \widetilde{\upsilon}.$$

Equivalent formulation:

$$\widetilde{\theta} = \operatorname*{argmax}_{\theta \in \varTheta} \breve{L}(\theta),$$

where

$$\breve{L}(\theta) = \max_{\substack{\upsilon \in \varUpsilon \\ \varPi_0 \upsilon = \theta}} L(\upsilon)$$

Fisher Theorem for pMLE $\quad \breve{D}_0(\widetilde{\theta} - \theta^*) \xrightarrow{w} \mathcal{N}(0, I_q),$

where

$$\breve{D}_0^2 = D_0^2 - A_0 H_0^{-2} A_0^\top,$$

$$\mathcal{D}_0^2 = -\nabla^2 \mathbb{E} L(\upsilon^*) = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}.$$

Bayesian setup

$$\upsilon \mid Y \propto \exp\{L(\upsilon)\} \pi(d\upsilon) \; - \; \text{full posterior},$$

$$\theta \mid Y \propto \int_H \exp\{L(\upsilon)\} \pi(d\upsilon) \; - \; \text{marginal posterior}.$$

Semiparametric BvM Theorem

$$\breve{D}_0(\theta - \widetilde{\theta}) \mid Y \xrightarrow{w} \mathcal{N}(0, I_q).$$

Motivating examples

- Partially linear regression:

$$Y_i = \theta^\top X_i + g(V_i) + \varepsilon_i,\ i = 1, \ldots, n.$$

- Generalized linear model:

$$Y_i \sim P_w,$$

where

- $\ell(y, w) = \log \frac{dP_w}{d\mu_0}(Y) = yw - d(w)$ for a convex function $d(w)$,
- $w = \theta^\top X + g(V)$.

Target: optimal projector on low dimensional subspace (vector θ).

Nuisance: function $g(\cdot)$.

The popular choices are uniform prior on θ and Gaussian process prior on $g(\cdot)$.

Outline

Finite dimensional nuisance

$$\upsilon = (\theta, \eta), \qquad dim(\eta) < \infty.$$

Aim: to bound the probability of $\{\theta \in \mathcal{A}\}$ given Y for $\mathcal{A} \subset \Theta$:

$$\begin{aligned}
\mathbb{P}\bigl(\theta \in \mathcal{A} \mid Y\bigr) &= \frac{\int \exp\bigl\{L(\upsilon)\bigr\} \mathbf{1}\bigl\{\theta \in \mathcal{A}\bigr\} d\pi(\upsilon)}{\int \exp\bigl\{L(\upsilon)\bigr\} d\pi(\upsilon)} \\
&= \underbrace{\frac{\int \exp\bigl\{L(\upsilon)\bigr\} \mathbf{1}\bigl\{\theta \in \mathcal{A}, \upsilon \in \Upsilon_0(\mathtt{r}_0)\bigr\} d\pi(\upsilon)}{\int \exp\bigl\{L(\upsilon)\bigr\} d\pi(\upsilon)}}_{as\ for\ the\ full\ parameter} \\
&+ \underbrace{\frac{\int \exp\bigl\{L(\upsilon)\bigr\} \mathbf{1}\bigl\{\theta \in \mathcal{A}, \upsilon \notin \Upsilon_0(\mathtt{r}_0)\bigr\} d\pi(\upsilon)}{\int \exp\bigl\{L(\upsilon)\bigr\} d\pi(\upsilon)}}_{tails\ of\ the\ posterior}
\end{aligned}$$

Consequently, there are additional error therms.

Identifiability condition

Case study: Let's suppose that

1. $\upsilon = (\theta, \eta) \sim \mathcal{N}(0, \mathcal{D}_0^{-2})$ — Gaussian random vector
2. $Corr(\theta, \eta) \simeq 1$

Then

$$\breve{D}_0^2 = D_0^2 - A_0 H_0^{-2} A_0^\top \simeq O,$$

$$\mathcal{D}_0^2 = \begin{pmatrix} D_0^2 & A_0 \\ A_0^\top & H_0^2 \end{pmatrix}.$$

and

$$\frac{\det(\mathcal{D}_0^2)}{(2\pi)^{p/2}} \int_{\mathbb{R}^{p-q}} \exp\left\{-\frac{1}{2}\|\mathcal{D}_0 \upsilon\|^2\right\} d\eta = \frac{\det(\breve{D}_0^2)}{(2\pi)^{q/2}} \exp\left\{-\frac{1}{2}\|\breve{D}_0 \eta\|^2\right\}$$

$\theta \sim \mathcal{N}(0, \breve{D}_0^{-2}) \simeq uniform(0) \Rightarrow$ no concentration!

Thus, the following condition is needed

$$(\mathbf{I}) \qquad \|D_0^{-1} A_0 H_0^{-2} A_0^\top D_0^{-1}\|_\infty \le \nu < 1.$$

Theorem

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

$$\|\breve{D}_0(\overline{\theta} - \theta^\circ)\|^2 \leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 16\mathrm{e}^{-\mathtt{x}}$$

$$\left\| I_q - \breve{D}_0 \mathfrak{S}^2 \breve{D}_0 \right\|_\infty \leq 4\Delta(\mathtt{r}_0, \mathtt{x}) + 16\mathrm{e}^{-\mathtt{x}}.$$

Moreover, for any measurable set $\mathcal{A} \subset \mathbb{R}^q$

$$\exp\bigl(-2\Delta(\mathtt{r}_0, \mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr) - \tau\Bigr\} - \mathrm{e}^{-\mathtt{x}}$$

$$\leq \mathbb{P}\Bigl(\mathfrak{S}^{-1}(\theta - \overline{\theta}) \in \mathcal{A} \Bigm| Y\Bigr)$$

$$\leq \exp\bigl(2\Delta(\mathtt{r}_0, \mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr) + \tau\Bigr\}.$$

where $\tau = \frac{1}{2}\sqrt{(1 + \Delta(\mathtt{r}_0, \mathtt{x}))^2 \Delta(\mathtt{r}_0, \mathtt{x})^2 + \Delta(\mathtt{r}_0, \mathtt{x})^2 q}$ and $\gamma \in \mathbb{R}^p$ is a standard Gaussian.

This shows that required condition for validity of results: "$\Delta(\mathtt{r}_0, \mathtt{x}) q^{1/2}$ is small".

Infinite dimensional nuisance

Consider (θ, f)-setup, where $\theta \in \varTheta \subseteq \mathbb{R}^q$ and $f \in \mathscr{H}$ for some separable Hilbert space $\mathscr{H}$. Let's suppose, that in $\mathscr{H}$ exists a countable basis (e_k). Then,

$$f = f(\eta) = \sum_{i=1}^{\infty} \eta_i e_i \in \mathscr{H}$$

where a vector $\eta = \{\eta_i\}_{i=1}^{\infty} \in l_2$ and $\eta_i = \langle f, e_i \rangle$.
Let the likelihood for full model is $\mathcal{L}(\theta, f(\eta))$. Denote

$$\mathcal{L}(\upsilon) = \mathcal{L}\bigl(\theta, f(\eta)\bigr) = \mathcal{L}\left(\theta, \sum_{i=1}^{\infty} \eta_i e_i\right),$$

where as usual $\upsilon = (\theta, \eta)$.

Infinite dimensional nuisance

If $\sum_{i=m+1}^{\infty} \eta_i^2$ is small, then

$$f(\eta) = \sum_{i=1}^{\infty} \eta_i e_i \simeq \sum_{i=1}^{m} \eta_i e_i$$

Corresponding finite dimensional approximation of log-likelihood:

$$\mathcal{L}_m(\upsilon_m) = \mathcal{L}(\theta, \sum_{i=1}^{m} \eta_i e_i),$$

where $\upsilon_m = (\theta, \eta_m)$.
Thus, we have finite dimensional versions of true-parameter, Fisher information and score function:

$$\upsilon_m^* \stackrel{\text{def}}{=} \operatorname*{argmax}_{\upsilon_m \in \Upsilon_m} \mathbb{E}\mathcal{L}_m(\upsilon_m), \quad \mathcal{D}_m^2 \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}\mathcal{L}_m(\upsilon_m^*), \quad \xi_m \stackrel{\text{def}}{=} \mathcal{D}_m^{-1} \nabla \mathcal{L}_m(\upsilon_m^*).$$

Bias condition

Two kinds of bias arise:

- $\theta^* - \theta_m^*$, where

$$\theta^* = \operatorname*{argmax}_{\theta \in \Theta} \max_{f \in \mathscr{H}} I\!E \mathcal{L}(\theta, f), \qquad \theta_m^* = \Pi_0 \upsilon_m^*.$$

- $\breve{D}_0^2 - \breve{D}_m^2$, where $\breve{D}_m^2 \stackrel{\text{def}}{=} \left(\Pi_0 \mathcal{D}_m^{-2} \Pi_0^\top\right)^{-1} \in I\!R^{q \times q}$,

$$\breve{D}_0^2 \stackrel{\text{def}}{=} \left(\Pi_0 \mathcal{D}_0^{-2} \Pi_0^\top\right)^{-1} \in I\!R^{q \times q}, \quad \mathcal{D}_0^2 \stackrel{\text{def}}{=} \nabla^2 I\!E[\mathcal{L}(\upsilon^*)] \in Lin(l_2, l_2),$$

and $Lin(l_2, l_2)$ is the space of linear operators from l_2 to l_2.

(B) For any $m \in \mathbb{N}$ there exist constants α_m and β_m such that

$$\|\breve{D}_0(\theta^* - \theta_m^*)\|^2 \leq \alpha_m,$$

$$\left\| I_q - \breve{D}_0 \breve{D}_m^{-2} \breve{D}_0 \right\|_\infty \leq \beta_m.$$

Sieve prior

Natural prior for sieve approximations:

- Consider non-informative prior on θ and sieve prior on η:
  $\pi^m(\eta) = \Pi_{i=1}^{\infty} \pi_i(\eta_k) = \Pi_{i=1}^{m} \pi_i(\eta_k)$, i.e. $\pi_i \equiv 0, i > m$.
- Thus, posterior depends only on θ and $\eta^m = \{\eta_k\}_{k=1}^m$.
- Then, under proper conditions we have a BvM result for marginal distribution of θ component of full parameter $\upsilon_m = (\theta, \eta_m)$.

Theorem

Define $\overline{\theta}_m \stackrel{\text{def}}{=} \mathbb{E}\bigl(\theta \,\big|\, Y\bigr)$, $\mathfrak{S}_m^2 \stackrel{\text{def}}{=} \operatorname{Cov}\bigl(\theta \,\big|\, Y\bigr)$ *and* $\theta_m^{\circ} = \theta_m^* + \breve{D}_m^{-1} \breve{\xi}_m$,
where $\breve{\xi}_m = \Pi_0 \xi_m = \Pi_0 \mathcal{D}_m^{-1} \nabla \mathcal{L}_m(\upsilon_m^*)$.
Then, under condition (B) *it holds with dominating probability (*$\geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$*):*

$$\|\breve{D}_0(\overline{\theta}_m - \theta_m^{\circ})\|^2 \leq (1+\beta_m)\Delta_{\circ} + \alpha_m.$$

$$\bigl\|\boldsymbol{I}_q - \breve{D}_0 \mathfrak{S}_m^2 \breve{D}_0\bigr\|_{\infty} \leq \beta_m + (1+\beta_m)\Delta_{\circ}.$$

Moreover, for any measurable set $\mathcal{A} \subset \mathbb{R}^q$

$$\exp\bigl(-\Delta_{\circ}\bigr)\Bigl\{\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr) - \frac{1}{2}\sqrt{(1+\Delta_{\circ})^2 \Delta_{\circ}^2 + \Delta_{\circ}^2 q}\Bigr\} - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$$
$$\leq \mathbb{P}\Bigl(\mathfrak{S}_m^{-1}(\theta - \overline{\theta}_m) \in \mathcal{A} \,\Big|\, Y\Bigr)$$
$$\leq \exp\bigl(\Delta_{\circ}\bigr)\Bigl\{\mathbb{P}\bigl(\gamma \in \mathcal{A}\bigr) + \frac{1}{2}\sqrt{(1+\Delta_{\circ})^2 \Delta_{\circ}^2 + \Delta_{\circ}^2 q}\Bigr\} + \mathtt{C}\mathrm{e}^{-\mathtt{x}}.$$

where $\Delta_{\circ} = \mathtt{C}\sqrt{\frac{(p_m + \mathtt{x})^3}{n}}$, $p_m = \dim(\theta) + m$ *and* $\gamma \in \mathbb{R}^q$ *is a standard Gaussian.*

Theorem

Define $\overline{\theta}_m \stackrel{\text{def}}{=} I\!E(\theta \,|\, Y)$, $\mathfrak{S}_m^2 \stackrel{\text{def}}{=} \operatorname{Cov}(\theta \,|\, Y)$ *and* $\theta_m^\circ = \theta_m^* + \breve{D}_m^{-1} \breve{\xi}_m$,
where $\breve{\xi}_m = \Pi_0 \xi_m = \Pi_0 \mathcal{D}_m^{-1} \nabla \mathcal{L}_m(\upsilon_m^*)$.
Then, under condition (B) *it holds with dominating probability* ($\geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$):

$$\|\breve{D}_0(\overline{\theta}_m - \theta_m^\circ)\|^2 \leq (1+\beta_m)\Delta_\circ + \alpha_m .$$

$$\left\| I_q - \breve{D}_0 \mathfrak{S}_m^2 \breve{D}_0 \right\|_\infty \leq \beta_m + (1+\beta_m)\Delta_\circ .$$

Moreover, for any measurable set $\mathcal{A} \subset I\!R^q$

$$\exp\bigl(-2\Delta(\mathtt{r}_0,\mathtt{x}) - 8\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{I\!P(\gamma \in \mathcal{A}) - \tau\Bigr\} - \mathrm{e}^{-\mathtt{x}}$$

$$\leq I\!P\bigl(\mathfrak{S}_m^{-1}(\theta - \overline{\theta}_m) \in \mathcal{A} \,\big|\, Y\bigr)$$

$$\leq \exp\bigl(2\Delta(\mathtt{r}_0,\mathtt{x}) + 5\mathrm{e}^{-\mathtt{x}}\bigr)\Bigl\{I\!P(\gamma \in \mathcal{A}) + \tau\Bigr\}.$$

where $\Delta_\circ = 4\Delta(\mathtt{r}_0,\mathtt{x}) + 16\mathrm{e}^{-\mathtt{x}}$ *and* $\gamma \in \mathbb{R}^q$ *is a standard Gaussian.*

Example: GLM

Generalized linear model: introduction

$$Y_i \sim P_w,$$

where

- $\ell(y,w) = \log \frac{dP_w}{d\mu_0}(Y) = yw - d(w)$ for a convex function $d(w)$,
- $w = \upsilon^\top X$.

The corresponding log-density of a GLM reads as

$$\mathcal{L}(\theta) = \sum \bigl\{ Y_i \Psi_i^\top \upsilon - d(\Psi_i^\top \upsilon) \bigr\}.$$

Define now the matrix $\mathcal{D}_0^2$ by

$$\mathcal{D}_0^2 \stackrel{\text{def}}{=} -\nabla^2 \mathcal{L}(\upsilon^*) = \sum d''(\Psi_i^\top \upsilon^*) \Psi_i \Psi_i^\top.$$

Example: conditions

Define the individual errors (residuals)

$$\varepsilon_i = Y_i - \mathbb{E} Y_i .$$

(e_1) *There exist some constants* v_0 *and* $\mathrm{g}_1 > 0$*, and for every* i *a constant* $\mathfrak{s}_i$ *such that* $\mathbb{E}\bigl(\varepsilon_i / \mathfrak{s}_i\bigr)^2 \le 1$ *and*

$$\log \mathbb{E} \exp\bigl(\mu \varepsilon_i / \mathfrak{s}_i\bigr) \le \mathrm{v}_0^2 \mu^2 / 2, \qquad |\mu| \le \mathrm{g}_1 .$$

It allows to check exponential moment conditions. In particular, $\omega \equiv 0$.

Example: conditions

Let $d''(z)$ be Lipschitz continuous:

$$|d''(z) - d''(z_0)| \le L|z - z_0|, \quad z, z_0 \in \mathbb{R}$$

Then

$$\delta(\mathtt{r}) \le L \frac{\mathtt{r}}{N_2^{1/2}},$$

where

$$N_2^{-1/2} \stackrel{\text{def}}{=} \max_i \sup_{\gamma \in \mathbb{R}^p} \frac{|\Psi_i^\top \gamma|}{d''(\Psi_i^\top \upsilon^*) \cdot \|\mathcal{D}_0 \gamma\|}.$$

Finally,

$$\Delta(\mathtt{r}_0, \mathtt{x}) \stackrel{\text{def}}{=} \left\{ \delta(\mathtt{r}_0) + 6\nu_0\, p^{1/2}\, \omega \right\} \mathtt{r}_0^2 \le L \frac{\mathtt{r}_0^3}{N_2^{1/2}} \sim \mathtt{C} \frac{p^{3/2}}{n^{1/2}}$$

Example: semiparametric case

Generalized linear model:

$$Y_i \sim P_w,$$

where

- $\ell(y,w) = \log \frac{dP_w}{d\mu_0}(Y) = yw - d(w)$ for a convex function $d(w)$,
- $w = \theta^\top X + \eta(V)$.

Sieve approach:

- Suppose $\eta \in \mathscr{F}_\beta = \{\eta(V) = \sum_{i=1}^{\infty} \eta_k \phi_k(V) \colon \sum_{k=1}^{\infty} \eta_k^2 k^{-2\beta} \le L\}$, where $\{\phi_i\}_{i=1}^{\infty}$ is some orthonormal functional basis (i.e. Fourier, wavelet, etc.).
- Then it can be shown, that bias due to sieve approach is small for $\beta > \frac{3}{2}$:

$$\|\breve{D}_0(\theta^* - \theta_m^*)\|^2 \le \mathtt{C}\,\frac{n}{m^{2s}},$$

$$\left\|\boldsymbol{I}_q - \breve{D}_0 \breve{D}_m^{-2} \breve{D}_0\right\|_\infty \le \mathtt{C}\,\frac{1}{n}.$$

Outline

Outlook

- Smooth priors in large dimension including Gaussian priors. The approach extends by considering penalized likelihood.

- Hyperpriors. The approach extends under smoothness of hyperpriors (*in work of Spokoiny and Baldin*).

- Time-series analysis. The approach can be extended by use of penalization technique (*in forthcoming paper by Spokoiny and P.*).

References

- Freedman (AoS 1999) BvM with infinite dimensional parameters.
- Ghosal (Bernoulli 1999). BvM in high-dimensional linear models.
- Ghosal (JoMA 2000). BvM for exponential families.
- Ghosal, Ghosh, and Van der Vaart (AoS 2000). Convergence rates of posterior distributions.
- Shen and Wasserman (AoS 2001). Rates of convergence of posterior distributions
- Ghosal, Ghosh, and Van der Vaart (AoS 2007). Convergence rates of posterior distributions for non i.i.d. observations.
- Boucheron and Gassiat (EJS 2009). BvM theorem for discrete probability distributions.
- Castillo (PTRF 2010). Semi-parametric BvM.
- Rivoirard and Rousseau (AoS 2012). BvM for linear functionals of the density.
- Bontemps (AoS 2011). BvM for Gaussian regression with increasing number of regressors.