Weierstraß-Institut für
Angewandte Analysis und Stochastik

Leibniz-Gemeinschaft

Fisher and Wilks expansions with applications to statistical inference

Vladimir Spokoiny,
WIAS, HU Berlin

Outline

Data $\boldsymbol{Y} \sim I\!P$. Aim: infer on $I\!P$.

Parametric assumption (PA): $I\!P \in \left(I\!P_{\boldsymbol{\theta}} \, , \, \boldsymbol{\theta} \in \Theta \subseteq I\!R^{p} \right) \ll \boldsymbol{\mu}_0$.

Maximum likelihood estimator (MLE):

$$\tilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \log \frac{dI\!P_{\boldsymbol{\theta}}}{d\boldsymbol{\mu}_0}(\boldsymbol{Y})$$

PA-PW: $I\!P \notin \left(I\!P_{\boldsymbol{\theta}} \right)$. Target of estimation ?

$$\boldsymbol{\theta}^{*} \stackrel{\text{def}}{=} \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} I\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} I\!E \log \frac{dI\!P_{\boldsymbol{\theta}}}{dI\!P} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{K}(I\!P, I\!P_{\boldsymbol{\theta}}).$$

Under PA: $I\!P = I\!P_{\boldsymbol{\theta}^{*}}$ and

$$\operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \mathcal{K}(I\!P_{\boldsymbol{\theta}^{*}}, I\!P_{\boldsymbol{\theta}}) = \boldsymbol{\theta}^{*}$$

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

Theorem

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

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

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

with

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

Here $\diamondsuit(\mathtt{x})$ *and* $\Delta(\mathtt{x})$ *are explicit error terms.*

A linear model

Given

- $\boldsymbol{Y}$, response,
- $\Sigma = \operatorname{Cov}(\boldsymbol{Y})$, its covariance matrix
- Ψ , design matrix of regressors:

$$\boldsymbol{Y} = \Psi^{\top}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \qquad \mathbb{E}\boldsymbol{\varepsilon} = 0, \qquad \operatorname{Cov}(\boldsymbol{\varepsilon}) = \Sigma.$$

PA: $\boldsymbol{Y} \sim \mathcal{N}(\Psi^{\top}\boldsymbol{\theta}, \Sigma)$:

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta})^{\top}\Sigma^{-1}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}) + R$$

Study under true: $\mathbb{E}\boldsymbol{Y} = \boldsymbol{f}$ and $\operatorname{Cov}(\boldsymbol{Y}) = \Sigma_0$.

$$L(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta})^{\top}\Sigma^{-1}(\boldsymbol{Y} - \Psi^{\top}\boldsymbol{\theta}) + R,$$

Lemma

$L(\boldsymbol{\theta})$ *is quadratic in* $\boldsymbol{\theta}$ *and it holds with* $\mathbb{E}\boldsymbol{Y} = \boldsymbol{f}$, $\boldsymbol{\varepsilon} \stackrel{\text{def}}{=} \boldsymbol{Y} - \boldsymbol{f}$ *:*

$$\begin{aligned}
\nabla^2 L(\boldsymbol{\theta}^*) &= -\Psi\Sigma^{-1}\Psi^{\top} \\
D_0^2 &\stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}^*) = \Psi\Sigma^{-1}\Psi^{\top}, \\
\widetilde{\boldsymbol{\theta}} &= D_0^{-2}\Psi\Sigma^{-1}\boldsymbol{Y}, \\
\boldsymbol{\theta}^* &= D_0^{-2}\Psi\Sigma^{-1}\boldsymbol{f}, \\
\boldsymbol{\xi} &\stackrel{\text{def}}{=} D_0^{-1}\nabla L(\boldsymbol{\theta}^*) = D_0^{-1}\Psi\Sigma^{-1}\boldsymbol{\varepsilon}, \\
D_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) &\equiv \boldsymbol{\xi}, \\
L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) &\equiv \|\boldsymbol{\xi}\|^2/2.
\end{aligned}$$

Fisher and Wilks for a linear model

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= -\frac{1}{2}(\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta})^\top \Sigma^{-1} (\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}) + R, \\
\nabla L(\boldsymbol{\theta}) &= \Psi \Sigma^{-1} (\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}), \\
\nabla^2 L(\boldsymbol{\theta}) &\equiv -\Psi \Sigma^{-1} \Psi^\top
\end{aligned}
$$

$L(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$ and it holds with $\mathbb{E}\boldsymbol{Y} = \boldsymbol{f}$, $\boldsymbol{\varepsilon} \stackrel{\text{def}}{=} \boldsymbol{Y} - \boldsymbol{f}$:

$$
\begin{aligned}
D_0^2 &\stackrel{\text{def}}{=} -\nabla^2 \mathbb{E} L(\boldsymbol{\theta}) & &= \Psi \Sigma^{-1} \Psi^\top, \\
\widetilde{\boldsymbol{\theta}} &= D_0^{-2} \Psi \Sigma^{-1} \boldsymbol{Y}, & \nabla L(\widetilde{\boldsymbol{\theta}}) &= \Psi \Sigma^{-1} (\boldsymbol{Y} - \Psi^\top \widetilde{\boldsymbol{\theta}}) = 0, \\
\boldsymbol{\theta}^* &= D_0^{-2} \Psi \Sigma^{-1} \boldsymbol{f}, & \nabla \mathbb{E} L(\boldsymbol{\theta}^*) &= \Psi \Sigma^{-1} (\boldsymbol{f} - \Psi^\top \boldsymbol{\theta}^*) = 0, \\
\boldsymbol{\xi} &\stackrel{\text{def}}{=} D_0^{-1} \nabla L(\boldsymbol{\theta}^*) & &= D_0^{-1} \Psi \Sigma^{-1} (\boldsymbol{Y} - \Psi^\top \boldsymbol{\theta}^*) = D_0^{-1} \Psi \Sigma^{-1} \boldsymbol{\varepsilon}.
\end{aligned}
$$

Hence $D_0 \bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) = \boldsymbol{\xi}$ and

$$
L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}^*) = -\frac{1}{2} (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\top \nabla^2 L(\widetilde{\boldsymbol{\theta}}) (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = -\frac{1}{2} \bigl\| D_0 \bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\bigr) \bigr\|^2 = -\frac{1}{2} \|\boldsymbol{\xi}\|^2,
$$

Fisher and Wilks for a linear model. Inference under PA

Under PA: $\boldsymbol{Y} \sim \mathcal{N}(\Psi^{\top}\boldsymbol{\theta}^{*}, \Sigma)$.

Then $\boldsymbol{\xi} = D_0^{-1}\Psi\Sigma^{-1}(\boldsymbol{Y} - \mathbb{E}\boldsymbol{Y})$ is normal zero mean and

$$\operatorname{Var}(\boldsymbol{\xi}) = \operatorname{Var}\bigl(D_0^{-1}\Psi\Sigma^{-1}\boldsymbol{\varepsilon}\bigr) = D_0^{-1}\Psi\Sigma^{-1}\operatorname{Var}(\boldsymbol{\varepsilon})\Sigma^{-1}\Psi D_0^{-1} = I_p\,.$$

Therefore, $D_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*}\bigr) = \boldsymbol{\xi}$ is standard normal and

$$2L(\widetilde{\boldsymbol{\theta}}) - 2L(\boldsymbol{\theta}^{*}) = \|\boldsymbol{\xi}\|^2 \sim \chi_p^2$$

If z_α^2 is the $1-\alpha$ quantile of χ_p^2, then

$$\mathcal{E}(z_\alpha) = \bigl\{\boldsymbol{\theta}\colon \|D_0\bigl(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}\bigr)\| \le z_\alpha\bigr\} = \bigl\{\boldsymbol{\theta}\colon L(\widetilde{\boldsymbol{\theta}}) - L(\boldsymbol{\theta}) \le z_\alpha^2/2\bigr\}$$

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

$$\mathbb{P}\bigl(\boldsymbol{\theta}^{*} \notin \mathcal{E}(z_\alpha)\bigr) = \alpha.$$

- The Fisher and Wilks expansions are only based on geometric features of the likelihood ($L(\boldsymbol{\theta})$ is quadratic in $\boldsymbol{\theta}$).
- The true distribution is not involved.
- Applies for any sample size.
- For inference, the PA is important. It only concerns the distribution of $\boldsymbol{\xi}$.
- PA-PW: Let $\operatorname{Var}\bigl(\boldsymbol{Y}\bigr) = \Sigma_0 \neq \Sigma$. Then with $D_0^2 = \Psi \Sigma^{-1} \Psi^\top$

$$
\operatorname{Var}\bigl\{\nabla L(\boldsymbol{\theta}^*)\bigr\} = \operatorname{Var}\bigl\{\Psi \Sigma^{-1} \boldsymbol{Y}\bigr\} = \Psi \Sigma^{-1} \Sigma_0 \Sigma^{-1} \Psi^\top \stackrel{\text{def}}{=} V_0^2 \neq D_0^2
$$

and (the sandwich formula)

$$
\operatorname{Var}(\boldsymbol{\xi}) = \operatorname{Var}\bigl\{D_0^{-1} \nabla L(\boldsymbol{\theta}^*)\bigr\} = D_0^{-1} V_0^2 D_0^{-1} \neq I_p .
$$

Ley $\boldsymbol{Y} = (Y_1, \ldots . Y_n)$ be i.i.d. from P.
PA: $P \in (P_{\boldsymbol{\theta}}\,,\ \boldsymbol{\theta} \in \Theta)$, a regular family with $\ell(y, \boldsymbol{\theta}) = \log p(y, \boldsymbol{\theta})$.

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\theta}), \qquad \widetilde{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}).$$

Theorem

Assume PA: $P = P_{\boldsymbol{\theta}^*} \in (P_{\boldsymbol{\theta}})$. Then

$$\sqrt{n\mathbb{F}_{\boldsymbol{\theta}^*}}\bigl(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\bigr) \xrightarrow{w} \mathcal{N}(0, \boldsymbol{I}_p),$$

$$L(\widetilde{\boldsymbol{\theta}}_n) - L(\boldsymbol{\theta}^*) \xrightarrow{w} \chi_p^2/2$$

where $\mathbb{F}_{\boldsymbol{\theta}^*}$ is the Fisher information matrix:

$$\mathbb{F}_{\boldsymbol{\theta}^*} = -\nabla^2 E\,\ell(Y_1, \boldsymbol{\theta}^*) = \operatorname{Var}\bigl\{\nabla \ell(Y_1, \boldsymbol{\theta}^*)\bigr\}.$$

(Non-asymptotic) expansions:

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

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

where

$$D_0^2 = D_n^2 = -n\nabla^2 E\,\ell(Y_1, \boldsymbol{\theta}^*) = n\mathbb{F}_{\boldsymbol{\theta}^*}$$

$$\boldsymbol{\xi} = \boldsymbol{\xi}_n = (n\mathbb{F}_{\boldsymbol{\theta}^*})^{-1/2} \sum_{i=1}^{n} \nabla\ell(Y_i, \boldsymbol{\theta}^*)$$

Under PA $\nabla\ell(Y_i, \boldsymbol{\theta}^*)$ are i.i.d. zero mean with $\operatorname{Var}\bigl\{\nabla\ell(Y_1, \boldsymbol{\theta}^*)\bigr\} = \mathbb{F}_{\boldsymbol{\theta}^*}$, and by CLT

$$\boldsymbol{\xi}_n \xrightarrow{w} \mathcal{N}(0, I_p)$$

For

$$\widetilde{\boldsymbol{\theta}}_n = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\theta}),$$

it holds with $D_n^2 = n\mathbb{F}_{\boldsymbol{\theta}^*}$

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

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

The error terms satisfy

$$\diamondsuit_n(\mathtt{x}) \leq \mathtt{C} \sqrt{\frac{(p + \mathtt{x})^2}{n}}, \qquad \Delta_n(\mathtt{x}) \leq \mathtt{C} \sqrt{\frac{(p + \mathtt{x})^3}{n}}.$$

and

$$\|\boldsymbol{\xi}_n\|^2 \leq p + \mathtt{C}\mathtt{x}.$$

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;

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

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

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

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

Some existing results

[Portnoy, 1984]: M-estimator i.i.d. or linear models:

– $p_n \log(p_n)/n \to 0$, consistency;

– $p_n^2 \log^2(p)/n \to 0$, asymptotic normality; (a counterexample for $p^2/n \to \infty$).

[Portnoy, 1988]: MLE for a GLM:

– $p_n^{3/2} \log(n)/n \to 0$, Wilks Theorem $p_n^{-1/2} L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - p_n^{1/2} \xrightarrow{w} \mathcal{N}(0,1)$;

Sieve estimation:

[Birgé and Massart, 1993], [Chen, 1993, 1997], [Van de Geer, 1993, van de Geer, 2002]; ...

Outline

Aim:

- minimal non-restrictive and natural conditions
- possibly sharp bounds
- all constants explicit, no asymptotic arguments
- model misspecification incorporated
- self-contained

Main steps

- Concentration and large deviations: for some $\mathtt{r}_0$

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \Theta_0(\mathtt{r}_0)\bigr) \le \mathrm{e}^{-\mathtt{x}},$$

where $\Theta_0(\mathtt{r}) \stackrel{\mathrm{def}}{=} \bigl\{\boldsymbol{\theta} \colon \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\bigr\}$.

- Local quadratic approximation of the expected log-likelihood:

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \frac{2\mathbb{E}L(\boldsymbol{\theta}^*) - 2\mathbb{E}L(\boldsymbol{\theta})}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \le \delta(\mathtt{r}).$$

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

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

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

Define

$$D^2(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\nabla^2 I\!\!E L(\boldsymbol{\theta}).$$

Then $D_0^2 = D_0^2(\boldsymbol{\theta}^*)$.

$(\mathcal{L}_0)$ *For each* $\mathtt{r} \le \mathtt{r}_0$ *, there is a constant* $\delta(\mathtt{r}) \le 1/2$ *such that it holds for any* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r}) = \{\boldsymbol{\theta} \in \Theta \colon \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}\}$ *:*

$$\left\| D_0^{-1} D_0^2(\boldsymbol{\theta}) D_0^{-1} - I_p \right\|_\infty \le \delta(\mathtt{r}).$$

By the second order Taylor expansion at $\boldsymbol{\theta}^*$ for any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$:

$$\begin{aligned}
\left| -2I\!\!E L(\boldsymbol{\theta}) + 2I\!\!E L(\boldsymbol{\theta}^*) - \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| &\le \delta(\mathtt{r})\mathtt{r}^2, \\
\left\| D_0^{-1} \{\nabla I\!\!E L(\boldsymbol{\theta}) - \nabla I\!\!E L(\boldsymbol{\theta}^*)\} + D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\| & \\
\le \left\| \left\{ I_p - D_0^{-1} D^2(\boldsymbol{\theta}^\circ) D_0^{-1} \right\} D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\| &\le \delta(\mathtt{r})\mathtt{r}.
\end{aligned}$$

Aim: To bound the error of the local constant approximation of the gradient (vector) process

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl\| D_0^{-1} \bigl\{ \nabla \zeta(\boldsymbol{\theta}) - \nabla \zeta(\boldsymbol{\theta}^*) \bigr\} \bigr\|$$

$(\boldsymbol{E\!D_2})$ *There exist a value* $\omega > 0$ *and for each* $\mathtt{r} > 0$, *a constant* $\mathtt{g}(\mathtt{r}) > 0$ *such that* $\zeta(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$ *satisfies for any* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *:*

$$\sup_{\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in I\!\!R^p} \log I\!\!E \exp\biggl\{ \frac{\lambda}{\omega} \, \frac{\boldsymbol{\gamma}_1^\top \nabla^2 \zeta(\boldsymbol{\theta}) \boldsymbol{\gamma}_2}{\|D_0 \boldsymbol{\gamma}_1\| \cdot \|D_0 \boldsymbol{\gamma}_2\|} \biggr\} \le \frac{\nu_0^2 \lambda^2}{2} , \qquad |\lambda| \le \mathtt{g}(\mathtt{r}).$$

Meaning: The second derivative of $\zeta(\boldsymbol{\theta})$ w.r.t. the local argument $\boldsymbol{\upsilon} = D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ is small.

Usually $\omega \asymp \|D_0^{-1}\| \asymp n^{-1/2}$.

Use $\boldsymbol{\upsilon} = D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)$ and consider $\mathcal{Y}(\boldsymbol{\upsilon}) = \omega^{-1} D_0^{-1}\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\}$:

$$\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\| D_0^{-1}\{\nabla\zeta(\boldsymbol{\theta}) - \nabla\zeta(\boldsymbol{\theta}^*)\}\bigr\| = \omega \sup_{\boldsymbol{\upsilon}\in\Upsilon_\circ(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\|,$$

$$\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon}\colon \|\boldsymbol{\upsilon}\| \le \mathtt{r}\}.$$

For any $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in I\!R^p$ with $\|\boldsymbol{\gamma}_1\| = \|\boldsymbol{\gamma}_2\| = 1$, condition (ED_2) implies

$$\log I\!\!E \exp\Bigl\{\lambda \boldsymbol{\gamma}_1^\top \nabla\mathcal{Y}(\boldsymbol{\upsilon})\boldsymbol{\gamma}_2\Bigr\} = \log I\!\!E \exp\Bigl\{\frac{\lambda}{\omega}\boldsymbol{\gamma}_1^\top D_0^{-1}\nabla^2\zeta(\boldsymbol{\theta})D_0^{-1}\boldsymbol{\gamma}_2\Bigr\} \le \frac{\nu_0^2\lambda^2}{2}.$$

A bound for the norm of a vector stochastic process

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

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

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:*

$$z_{\mathbb{H}}(\mathtt{x}) = \begin{cases} \sqrt{\mathbb{H}_2 + 2\mathtt{x}}, & \text{if } \mathbb{H}_2 + 2\mathtt{x} \le \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}\bigl(\mathtt{g}^{-1}\mathbb{H}_2 + \mathtt{g}\bigr), & \text{if } \mathbb{H}_2 + 2\mathtt{x} > \mathtt{g}^2. \end{cases}$$

Here $\mathbb{H}_2 = 4p$ *and* $\mathbb{H}_1 = 2p^{1/2}$, $\mathtt{g} = \mathtt{g}(\mathtt{r})$.

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 $z_{\mathbb{H}}(\mathtt{x})$.

Define

$$\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \stackrel{\text{def}}{=} D_0^{-1}\bigl\{\nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^*) + D_0^2\,(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr\}.$$

By Theorem 5

$$\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r}_0)} \bigl\|\chi(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\| \le \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

Suppose that $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$ on $\Omega(\mathtt{x})$. Then

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

The use of $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ yields the Fisher expansion.

A quadratic approximation

Define $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) = (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ})^{\top} \nabla L(\boldsymbol{\theta}^{\circ}) - \| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}) \|^2 / 2$ and

$$
\begin{aligned}
\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) & \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^{\circ}) - (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ})^{\top} \nabla L(\boldsymbol{\theta}^{\circ}) + \frac{1}{2} \| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}) \|^2 \\
& = L(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}), \qquad \boldsymbol{\theta}, \boldsymbol{\theta}^{\circ} \in \Theta_0(\mathtt{r})
\end{aligned}
$$

With $\boldsymbol{\theta}^{\circ}$ fixed, the gradient $\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) \stackrel{\text{def}}{=} \frac{d}{d\boldsymbol{\theta}} \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ})$ fulfills

$$
\nabla \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) = \nabla L(\boldsymbol{\theta}) - \nabla L(\boldsymbol{\theta}^{\circ}) + D_0^2 (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}) = D_0 \, \chi(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ});
$$

This implies

$$
\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) = (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ})^{\top} \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{\circ}),
$$

where $\boldsymbol{\theta}'$ is a point on the line connecting $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{\circ}$ and

$$
\bigl| \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^{\circ}) \bigr| = \bigl| (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ})^{\top} D_0 D_0^{-1} \nabla \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^{\circ}) \bigr| \le \| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^{\circ}) \| \sup_{\boldsymbol{\theta}' \in \Theta_0(\mathtt{r})} \bigl| \chi(\boldsymbol{\theta}', \boldsymbol{\theta}^{\circ}) \bigr| .
$$

$$
\begin{aligned}
\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}) &\stackrel{\mathrm{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^{\circ}) - (\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ})^{\top}\nabla L(\boldsymbol{\theta}^{\circ}) + \frac{1}{2}\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^{\circ})\|^2 \\
&= L(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}) - \mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}), \qquad \boldsymbol{\theta},\boldsymbol{\theta}^{\circ} \in \Theta_0(\mathtt{r})
\end{aligned}
$$

Theorem

Suppose $(\mathcal{L}_0)$ *,* (ED_0) *, and* (ED_2) *. For each* $\mathtt{r}$ *, it holds on a random set* $\Omega(\mathtt{r},\mathtt{x})$ *of a dominating probability at least* $1-\mathrm{e}^{-\mathtt{x}}$ *, it holds with any* $\boldsymbol{\theta},\boldsymbol{\theta}^{\circ}\in\Theta_0(\mathtt{r})$

$$
\begin{aligned}
\frac{\bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr|}{\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})\|} &\le \diamondsuit(\mathtt{r},\mathtt{x}), \qquad \bigl|\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^{*})\bigr| \le \mathtt{r}\,\diamondsuit(\mathtt{r},\mathtt{x}), \\
\frac{\bigl|\alpha(\boldsymbol{\theta}^{*},\boldsymbol{\theta})\bigr|}{\|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^{*})\|} &\le \diamondsuit(\mathtt{r},\mathtt{x}), \qquad \bigl|\alpha(\boldsymbol{\theta}^{*},\boldsymbol{\theta})\bigr| \le \mathtt{r}\,\diamondsuit(\mathtt{r},\mathtt{x}).
\end{aligned}
$$

Let $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$ on $\Omega(\mathrm{x})$. For $\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) = (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) - \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2/2$

$$\bigl|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\bigr| = \bigl|L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\bigr| \le \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathrm{x}), \qquad \boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r}_0)$$

The special case with $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^\circ = \widetilde{\boldsymbol{\theta}}$ yields in view of $\nabla L(\widetilde{\boldsymbol{\theta}}) = 0$ for $\mathtt{r} = \mathtt{r}_0$

$$\Bigl|L(\boldsymbol{\theta}^*) - L(\widetilde{\boldsymbol{\theta}}) + \|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2/2\Bigr| = \bigl|\alpha(\boldsymbol{\theta}^*, \widetilde{\boldsymbol{\theta}})\bigr| \le \mathtt{r}_0 \diamondsuit(\mathtt{r}_0, \mathrm{x}). \tag{1}$$

Further, on the set of a dominating probability, it holds $\|\boldsymbol{\xi}\| \le z(B, \mathrm{x})$ (later). Now

$$\begin{aligned}
&\bigl|\|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\|^2 - \|\boldsymbol{\xi}\|^2\bigr| \\
&\quad \le 2\,\|\boldsymbol{\xi}\| \cdot \|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\| + \|D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - \boldsymbol{\xi}\|^2 \\
&\quad \le 2\, z(B, \mathrm{x}) \diamondsuit(\mathtt{r}_0, \mathrm{x}) + \diamondsuit^2(\mathtt{r}_0, \mathrm{x}).
\end{aligned}$$

Together with (1), this yields

$$\bigl|L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \|\boldsymbol{\xi}\|^2/2\bigr| \le \bigl\{\mathtt{r}_0 + z(B, \mathrm{x})\bigr\} \diamondsuit(\mathtt{r}_0, \mathrm{x}) + \diamondsuit^2(\mathtt{r}_0, \mathrm{x})/2.$$

The error term can be improved if the squared root of the excess is considered.

Indeed, if $\widetilde{\boldsymbol{\theta}} \in \Theta_0(\mathtt{r}_0)$

$$\Bigl| \bigl\{ 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr\}^{1/2} - \| D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \| \Bigr| \leq \frac{\bigl| 2L(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) - \| D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|^2 \bigr|}{\| D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|}$$

$$\leq \frac{2 \bigl| \alpha(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}^*) \bigr|}{\| D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \|} \leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} \frac{2 \bigl| \alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \bigr|}{\| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \|} \leq 2 \diamondsuit(\mathtt{r}_0, \mathtt{x}).$$

The Fisher expansion allows to replace here the norm of the standardized error $D_0(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ with the norm of the normalized score $\boldsymbol{\xi}$.

Aim: find $\mathtt{r}_0$ ensuring

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

- By definition $\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r}_0)} L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \ge 0$. Suffices to check that

$$L(\boldsymbol{\theta},\boldsymbol{\theta}^*) < 0 \qquad \forall \boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)$$

- Use the decomposition

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

- Bound $\|\boldsymbol{\xi}\| = \|D_0^{-1}\nabla\zeta(\boldsymbol{\theta}^*)\|$;
- Upper function device for the remainder

$$\sup_{\boldsymbol{\theta}\in\Theta\setminus\Theta_0(\mathtt{r}_0)} \bigl\{\zeta(\boldsymbol{\theta},\boldsymbol{\theta}^*) - (\boldsymbol{\theta}-\boldsymbol{\theta}^*)^\top \nabla\zeta(\boldsymbol{\theta}^*) - f(\boldsymbol{\theta},\boldsymbol{\theta}^*)\bigr\} \le 0 \qquad \text{w.h.p.}$$

$(\mathcal{L})$ *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*

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

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

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

Then

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

The radius $\mathtt{r}_0$ has to fulfill

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

One can use that

- $\mathtt{b}(\mathtt{r}_0) \geq 1 - \delta(\mathtt{r}_0) \approx 1$,
- the constant ω and thus, $\varrho(\mathtt{r},\mathtt{x})$, is small, and
- $\mathtt{r}\mathtt{b}(\mathtt{r})$ grows with $\mathtt{r}$.

A simple rule $\mathtt{r}_0 \geq (2+\delta)z(B,\mathtt{x})$ for some $\delta > 0$ works in most of cases.

$(\boldsymbol{ED_0})$ *There exist a positive symmetric matrix* V_0^2 *, and constants* $\mathtt{g} > 0$ *,* $\nu_0 \ge 1$ *such that* $\operatorname{Var}\bigl\{ \nabla \zeta(\boldsymbol{\theta}^*) \bigr\} \le V_0^2$ *and*

$$\log I\!E \exp\bigl(\boldsymbol{\gamma}^\top V_0^{-1} \nabla \zeta(\boldsymbol{\theta}^*) \bigr) \le \frac{\nu_0^2 \|\boldsymbol{\gamma}\|^2}{2}, \qquad \boldsymbol{\gamma} \in I\!R^p, \ \|\boldsymbol{\gamma}\| \le \mathtt{g}.$$

With $\boldsymbol{\eta} = V_0^{1} \nabla \zeta(\boldsymbol{\theta}^*)$, it holds $\boldsymbol{\xi} = D_0^{-1} V_0 \boldsymbol{\eta}$ and

$$\|\boldsymbol{\xi}\|^2 = \boldsymbol{\eta}^\top B \boldsymbol{\eta}$$

for $B = D_0^{-1} V_0^2 D_0^{-1}$. Also define

$$\mathtt{p}_B \stackrel{\text{def}}{=} \operatorname{tr}\bigl(B\bigr), \qquad \mathtt{v}_B^2 \stackrel{\text{def}}{=} 2 \operatorname{tr}(B^2), \qquad \lambda_B \stackrel{\text{def}}{=} \lambda_{\max}\bigl(B\bigr).$$

Note that $\mathtt{p}_B = I\!E \|\boldsymbol{\xi}\|^2$. Moreover, if $\boldsymbol{\xi}$ is a Gaussian vector then $\mathtt{v}_B^2 = \operatorname{Var}\bigl(\|\boldsymbol{\xi}\|^2\bigr)$. If $V_0^2 = D_0^2$, then $\lambda_B = 1$.

Define $\mu_c = 2/3$, $\mathtt{p}_B = \operatorname{tr}(B)$, $\mathtt{v}_B^2 = 2\operatorname{tr}(B^2)$, and $\lambda_B = \lambda_{\max}(B)$

$$\mathtt{g}_c \stackrel{\text{def}}{=} \sqrt{\mathtt{g}^2 - \mu_c \mathtt{p}_B},$$

$$2\mathtt{x}_c \stackrel{\text{def}}{=} (\mathtt{g}^2/\mu_c - \mathtt{p}_B)/\lambda_B + \log\det\bigl(I_p - \mu_c B/\lambda_B\bigr). \tag{3}$$

Theorem (SP2012)

Let (ED_0) *hold with* $\nu_0 = 1$ *and* $\mathtt{g}^2 \ge 2\mathtt{p}_B$. *Then for each* $\mathtt{x} > 0$

$$I\!P\bigl(\|\boldsymbol{\xi}\| \ge z(B,\mathtt{x})\bigr) = I\!P\bigl(\|B^{1/2}\boldsymbol{\eta}\| \ge z(B,\mathtt{x})\bigr) \le 2e^{-\mathtt{x}} + 8.4e^{-\mathtt{x}_c},$$

where $z(B,\mathtt{x})$ *is defined with* $\mathtt{y}_c^2 \le \mathtt{p}_B + 6\lambda_B\mathtt{x}_c$ *by*

$$z^2(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_B + 2\mathtt{v}_B\mathtt{x}^{1/2}, & \mathtt{x} \le \mathtt{v}_B/(18\lambda_B), \\ \mathtt{p}_B + 6\lambda_B\mathtt{x}, & \mathtt{v}_B/(18\lambda_B) < \mathtt{x} \le \mathtt{x}_c, \\ \bigl|\mathtt{y}_c + 2\lambda_B(\mathtt{x} - \mathtt{x}_c)/\mathtt{g}_c\bigr|^2, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

$$\mathtt{p}_B = \operatorname{tr}(B), \qquad \mathtt{v}_B^2 = 2\operatorname{tr}(B^2), \qquad \lambda_B = \lambda_{\max}(B).$$

$$z^2(B,\mathtt{x}) \stackrel{\text{def}}{=} \begin{cases} \mathtt{p}_B + 2\mathtt{v}_B\mathtt{x}^{1/2}, & \mathtt{x} \le \mathtt{v}_B/(18\lambda_B), \\ \mathtt{p}_B + 6\lambda_B\mathtt{x}, & \mathtt{v}_B/(18\lambda_B) < \mathtt{x} \le \mathtt{x}_c, \\ \left|\mathtt{y}_c + 2\lambda_B(\mathtt{x}-\mathtt{x}_c)/\mathtt{g}_c\right|^2, & \mathtt{x} > \mathtt{x}_c. \end{cases}$$

Depending on the value $\mathtt{x}$, we observe three types of tail behavior of the quadratic form $\|\xi\|^2$:

- The sub-Gaussian regime for $\mathtt{x} \le \mathtt{v}_B/(18\lambda_B)$
- The Poissonian regime for $\mathtt{x} \le \mathtt{x}_c$
- The value $\mathtt{x}_c$ from (3) is of order $\mathtt{g}^2$. In all our results we suppose that $\mathtt{g}^2$ and hence, $\mathtt{x}_c$ is sufficiently large;

The quadratic form $\|\xi\|^2$ can be bounded with a dominating probability by $\mathtt{p}_B + 6\lambda_B\mathtt{x}$ for a proper $\mathtt{x}$.

A “squared norm” trick

Let $\boldsymbol{\xi}$ be a random vector in $I\!\!R^p$ satisfying the condition

$$\log I\!\!E \exp\bigl(\boldsymbol{\gamma}^\top \boldsymbol{\xi}\bigr) \le \frac{\nu_0^2 \|\boldsymbol{\gamma}\|^2}{2}, \qquad \boldsymbol{\gamma} \in I\!\!R^p,\ \|\boldsymbol{\gamma}\| \le \mathrm{g}.$$

For simplicity we take here $B = 1$.

Aim: to bound $\|\boldsymbol{\xi}\|^2$.

A sup-representation:

$$\|\boldsymbol{\xi}\|^2 = \sup_{\boldsymbol{\gamma} \in I\!\!R^p} \bigl\{ \boldsymbol{\gamma}^\top \boldsymbol{\xi} - \|\boldsymbol{\gamma}\|^2/2 \bigr\}, \qquad \|\boldsymbol{\xi}\| = \sup_{\boldsymbol{\gamma} \in I\!\!R^p :\ \|\boldsymbol{\gamma}\| \le 1} \boldsymbol{\gamma}^\top \boldsymbol{\xi}.$$

Too rough to get a sharp bound on $\|\boldsymbol{\xi}\|$ with entropy arguments.

An exp-representation: for any $\mu < 1$

$$\exp\bigl\{\mu \|\boldsymbol{\xi}\|^2/2\bigr\} = \mathtt{C}_p(\mu) \int_{I\!\!R^p} \exp\bigl\{\boldsymbol{\gamma}^\top \boldsymbol{\xi} - \|\boldsymbol{\gamma}\|^2/(2\mu)\bigr\} d\boldsymbol{\gamma}$$

The proof is based on the following bound: for each $\mathtt{r}$

$$I\!P\biggl(\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl| \zeta(\boldsymbol{\theta}) - \zeta(\boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*) \bigr| \geq 3 \nu_0 \, z_{\mathbb{H}}(\mathtt{x}) \, \omega \, \mathtt{r} \biggr) \leq \mathrm{e}^{-\mathtt{x}}.$$

This bound is a special case of the general result from Theorem 9 below. It implies by Theorem 10 with $\rho = 1/2$ on a set $\Omega(\mathtt{x})$ of probability at least $1 - \mathrm{e}^{-\mathtt{x}}$ that for all $\mathtt{r} \geq \mathtt{r}_0$ and all $\boldsymbol{\theta}$ with $\| D_0 (\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \leq \mathtt{r}$

$$\bigl| \zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla \zeta(\boldsymbol{\theta}^*) \bigr| \leq \varrho(\mathtt{r}, \mathtt{x}) \, \mathtt{r},$$

where

$$\varrho(\mathtt{r}, \mathtt{x}) = 6 \nu_0 \, z_{\mathbb{H}} \bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0) \bigr) \, \omega \, . \tag{4}$$

The use of $\nabla I\!E L(\boldsymbol{\theta}^*) = 0$ yields

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \bigl| L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - I\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) \bigr| \leq \varrho(\mathtt{r}, \mathtt{x}) \, \mathtt{r}.$$

By definition $\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq 0$. So, it suffices to check that $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \in \Theta \setminus \Theta_0(\mathtt{r}_0)$.

We know

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) \right| \leq \varrho(\mathtt{r}, \mathtt{x})\, \mathtt{r}.$$

Also $\|\boldsymbol{\xi}\| \leq z(B, \mathtt{x})$ on $\Omega(\mathtt{x})$ and for each $\mathtt{r} \geq \mathtt{r}_0$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \left| (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top \nabla L(\boldsymbol{\theta}^*) \right|$$

$$\leq \sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| \times \| D_0^{-1} \nabla L(\boldsymbol{\theta}^*) \| = \mathtt{r} \|\boldsymbol{\xi}\| \leq z(B, \mathtt{x})\, \mathtt{r}.$$

Condition $(\mathcal{L})$ implies $-2 I\!\!E L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \geq \mathtt{r}^2 \mathtt{b}(\mathtt{r})$ for each $\boldsymbol{\theta}$ with $\| D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \| = \mathtt{r}$. We conclude that the condition

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

ensure $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) < 0$ for all $\boldsymbol{\theta} \notin \Theta_0(\mathtt{r}_0)$ with a dominating probability.

Let $\mathcal{U}(\boldsymbol{\upsilon})$ be a smooth stochastic process on an open subset $\Upsilon \subseteq I\!R^p$, and $I\!E\mathcal{U}(\boldsymbol{\upsilon}) \equiv 0$.

$(\mathcal{E}D)$ *There exist* $\mathtt{g} > 0$, $\nu_0 \ge 1$, *and a symmetric* $H_0 \ge 0$ *s.t. it holds*

$$\sup_{\boldsymbol{\gamma} \in I\!R^p :\, \|\boldsymbol{\gamma}\|=1} \log I\!E \exp\Bigl\{\lambda \frac{\boldsymbol{\gamma}^\top \nabla \mathcal{U}(\boldsymbol{\upsilon})}{\|H_0 \gamma\|}\Bigr\} \le \frac{\nu_0^2 \lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}.$$

We consider the local sets of the elliptic form $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0)\| \le \mathtt{r}\}$.

Theorem

Let $(\mathcal{E}D)$ *hold with some* $\mathtt{g} > 0$, *and a matrix* H_0. *For any* $\mathtt{x} \ge 0$ *and any* $\mathtt{r} > 0$

$$I\!P\Bigl\{\sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \bigl|\mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}_0)\bigr| \ge 3\nu_0\, \mathtt{r}\, z_{\mathbb{H}}(\mathtt{x})\Bigr\} \le \mathrm{e}^{-\mathtt{x}},$$

where $z_{\mathbb{H}}(\mathtt{x})$ *is given by the following rule: with* $\mathbb{H} = 4p$

$$z_{\mathbb{H}}(\mathtt{x}) = \begin{cases} \sqrt{\mathbb{H} + 2\mathtt{x}} & \text{if } \mathbb{H} + 2\mathtt{x} \le \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}\bigl(\mathtt{g}^{-1}\mathbb{H} + \mathtt{g}\bigr) & \text{if } \mathbb{H} + 2\mathtt{x} > \mathtt{g}^2, \end{cases}$$

On $\Omega(\mathtt{r}, \mathtt{x})$, one can bound $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) \stackrel{\text{def}}{=} \mathcal{U}(\boldsymbol{\upsilon}) - \mathcal{U}(\boldsymbol{\upsilon}_0)$:

$$\bigl|\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0)\bigr| \le 3\nu_0 \, \mathtt{r} \, z_{\mathbb{H}}(\mathtt{x}).$$

Aim: to build an upper function $f(\cdot)$ s.t. $\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - f(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0)$ is bounded uniformly in all $\boldsymbol{\upsilon}$.

Theorem

Let $(\mathcal{E}D)$ *hold on* $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$. *Given* $\mathtt{r}_0 < \mathtt{r}^*$, *define* $f(\mathtt{r}, \mathtt{r}_0)$ *for some* $\rho < 1$ *as*

$$f(\mathtt{r}, \mathtt{r}_0) = 3\nu_0 \mathtt{r} \, z_{\mathbb{H}}\bigl(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\bigr), \quad \mathtt{r}_0 \le \mathtt{r} \le \mathtt{r}^*. \tag{5}$$

Then it holds

$$I\!\!P\biggl(\sup_{\mathtt{r}_0 \le \mathtt{r} \le \mathtt{r}^*} \sup_{\boldsymbol{\upsilon} \in \Upsilon_\circ(\mathtt{r})} \bigl\{\mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}_0) - f\bigl(\rho^{-1}\mathtt{r}, \mathtt{r}_0\bigr)\bigr\} \ge 0\biggr) \le \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}}.$$

If $\mathtt{g} = \infty$, then $z_{\mathbb{H}}(\mathtt{x}) = \sqrt{2\mathtt{x} + 4p}$ and ($\rho = 1/2$)

$$f(\mathtt{r}, \mathtt{r}_0) = 3\nu_0 \mathtt{r} \sqrt{2\mathtt{x} + 4p + 2\log(\mathtt{r}/\mathtt{r}_0)}.$$

Idea: split $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)$ into slices $\mathcal{B}_{\mathtt{r}_k}(\boldsymbol{\upsilon}_0) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\boldsymbol{\upsilon}_0)$ and apply Theorem 9 to each slice. By (5) and Theorem 9 for any $\mathtt{r} > \mathtt{r}_0$

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\upsilon}\in\mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)\setminus\mathcal{B}_{\rho\mathtt{r}}(\boldsymbol{\upsilon}_0)} \bigl\{ \mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - f\bigl(\mathtt{r},\mathtt{r}_0\bigr) \bigr\} \ge 0 \Bigr)$$

$$\le \mathbb{P}\Bigl(\frac{1}{3\nu_0\mathtt{r}} \sup_{\boldsymbol{\upsilon}\in\mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}_0)} \mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) \ge z_{\mathbb{H}}\bigl(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\bigr) \Bigr) \le \frac{\mathtt{r}_0}{\mathtt{r}} \mathrm{e}^{-\mathtt{x}}. \tag{6}$$

Define $\mathtt{r}_k = \mathtt{r}_0\rho^{-k}$ for $k = 0,1,2,\ldots$ and $k^* \stackrel{\mathrm{def}}{=} \log(\mathtt{r}^*/\mathtt{r}_0) + 1$. By (6)

$$\mathbb{P}\Bigl(\sup_{\boldsymbol{\upsilon}\in\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}_0)\setminus\mathcal{B}_{\mathtt{r}_0}(\boldsymbol{\upsilon}_0)} \Bigl\{ \mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - f\bigl(\rho^{-1} d(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0),\mathtt{r}_0\bigr) \Bigr\} \ge 0 \Bigr)$$

$$\le \sum_{k=1}^{k^*} \mathbb{P}\Bigl(\frac{1}{\mathtt{r}_k} \sup_{\boldsymbol{\upsilon}\in\mathcal{B}_{\mathtt{r}_k}(\boldsymbol{\upsilon}_0)\setminus\mathcal{B}_{\mathtt{r}_{k-1}}(\boldsymbol{\upsilon}_0)} \Bigl\{ \mathcal{U}(\boldsymbol{\upsilon},\boldsymbol{\upsilon}_0) - f\bigl(\mathtt{r}_k,\mathtt{r}_0\bigr) \Bigr\} \ge 0 \Bigr)$$

$$\le \mathrm{e}^{-\mathtt{x}} \sum_{k=1}^{k^*} \rho^k \le \frac{\rho}{1-\rho} \mathrm{e}^{-\mathtt{x}}.$$

Let $\mathcal{Y}(\boldsymbol{\upsilon})$, $\boldsymbol{\upsilon} \in \varUpsilon$, be a smooth centered random vector process with values in $I\!R^q$, where $\varUpsilon \subseteq I\!R^p$. Let also $\mathcal{Y}(\boldsymbol{\upsilon}_0) = 0$ for a fixed point $\boldsymbol{\upsilon}_0 \in \varUpsilon$. (w.l.g. $\boldsymbol{\upsilon}_0 = 0$).

Suppose that $\mathcal{Y}(\boldsymbol{\upsilon})$ satisfies for each $\boldsymbol{\gamma} \in I\!R^p$ and $\boldsymbol{\alpha} \in I\!R^q$ with $\|\boldsymbol{\gamma}\| = \|\boldsymbol{\alpha}\| = 1$

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \log I\!E \exp\Bigl\{\lambda \boldsymbol{\gamma}^{\top} \nabla \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{\alpha}\Bigr\} \le \frac{\nu_0^2 \lambda^2}{2}, \qquad \lambda^2 \le 2\mathrm{g}^2. \tag{7}$$

We aim to bound the maximum of the norm $\|\mathcal{Y}(\boldsymbol{\upsilon})\|$ over a ball

$$\varUpsilon_{\circ}(\mathtt{r}) = \bigl\{\boldsymbol{\upsilon} \in \varUpsilon \colon \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0\| \le \mathtt{r}\bigr\}.$$

Condition (7) implies for any $\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})$ with $\|\boldsymbol{\upsilon}\| \le \mathtt{r}$ and $\|\boldsymbol{\gamma}\| = 1$ in view of $\mathcal{Y}(\boldsymbol{\upsilon}_0) = 0$

$$\log I\!E \exp\Bigl\{\frac{\lambda}{\mathtt{r}} \boldsymbol{\gamma}^{\top} \mathcal{Y}(\boldsymbol{\upsilon})\Bigr\} \le \frac{\nu_0^2 \lambda^2 \|\boldsymbol{\upsilon}\|^2}{2\mathtt{r}^2}, \qquad \lambda^2 \le 2\mathrm{g}^2; \tag{8}$$

Use the representation

$$\|\mathcal{Y}(\boldsymbol{\upsilon})\| = \sup_{\|\boldsymbol{u}\| \le \mathtt{r}} \frac{1}{\mathtt{r}} \boldsymbol{u}^{\top} \mathcal{Y}(\boldsymbol{\upsilon}).$$

This implies for $\varUpsilon_{\circ}(\mathtt{r}) = \bigl\{ \boldsymbol{\upsilon} \in \varUpsilon \colon \|\boldsymbol{\upsilon} - \boldsymbol{\upsilon}_0\| \le \mathtt{r} \bigr\}$

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})} \|\mathcal{Y}(\boldsymbol{\upsilon})\| = \sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})} \sup_{\|\boldsymbol{u}\| \le \mathtt{r}} \frac{1}{\mathtt{r}} \boldsymbol{u}^{\top} \mathcal{Y}(\boldsymbol{\upsilon}).$$

Consider a bivariate process $\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v})$ of $\boldsymbol{u} \in I\!R^{q}$ and $\boldsymbol{v} \in \Upsilon \subset I\!R^{p}$.
By definition $I\!E\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v}) = 0$. Further, for $\boldsymbol{\gamma} = \boldsymbol{u}/\|\boldsymbol{u}\|$

$$\nabla_{\boldsymbol{u}}\bigl[\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v})\bigr] = \mathcal{Y}(\boldsymbol{v}), \qquad \nabla_{\boldsymbol{v}}\bigl[\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v})\bigr] = \boldsymbol{u}^{\top}\nabla\mathcal{Y}(\boldsymbol{v}) = \|\boldsymbol{u}\|\boldsymbol{\gamma}^{\top}\nabla\mathcal{Y}(\boldsymbol{v})$$

Suppose that $\boldsymbol{u} \in I\!R^{q}$ and $\boldsymbol{v} \in \Upsilon$ are such that $\|\boldsymbol{u}\|^2 + \|\boldsymbol{v}\|^2 \le 2\mathtt{r}^2$. By the Hölder inequality, (8), and (7), it holds for $\|\boldsymbol{\gamma}\| = \|\boldsymbol{\alpha}\| = 1$ and $\boldsymbol{v} \in \Upsilon_{\circ}(\mathtt{r})$

$$\begin{aligned}
&\log I\!E\exp\Bigl\{\frac{\lambda}{2\mathtt{r}}(\boldsymbol{\gamma},\boldsymbol{\alpha})^{\top}\nabla\bigl[\boldsymbol{u}^{\top}\mathcal{Y}(\boldsymbol{v})\bigr]\Bigr\} \\
&\qquad \le \frac{1}{2}\log I\!E\exp\Bigl\{\frac{\lambda}{\mathtt{r}}\boldsymbol{\gamma}^{\top}\mathcal{Y}(\boldsymbol{v})\Bigr\} + \frac{1}{2}\log I\!E\exp\Bigl\{\frac{\lambda}{\mathtt{r}}\boldsymbol{u}^{\top}\nabla\mathcal{Y}(\boldsymbol{v})\boldsymbol{\alpha}\Bigr\} \\
&\qquad \le \frac{1}{2}\log I\!E\exp\Bigl\{\frac{\lambda}{\mathtt{r}}\boldsymbol{\gamma}^{\top}\mathcal{Y}(\boldsymbol{v})\Bigr\} + \frac{1}{2}\log I\!E\exp\Bigl\{\frac{\lambda}{\mathtt{r}}\|\boldsymbol{u}\|\boldsymbol{\gamma}^{\top}\nabla\mathcal{Y}(\boldsymbol{v})\boldsymbol{\alpha}\Bigr\} \\
&\qquad \le \frac{\nu_0^2\lambda^2}{4\mathtt{r}^2}\bigl(\|\boldsymbol{v}\|^2 + \|\boldsymbol{u}\|^2\bigr) \le \frac{\nu_0^2\lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}.
\end{aligned}$$

Theorem

Let a random p *-vector process* $\mathcal{Y}(\boldsymbol{\upsilon})$ *for* $\boldsymbol{\upsilon} \in \varUpsilon \subseteq I\!\!R^p$ *fulfill* $\mathcal{Y}(\boldsymbol{\upsilon}_0) = 0$ *,* $I\!\!E \mathcal{Y}(\boldsymbol{\upsilon}) \equiv 0$ *, and the condition* (7) *be satisfied. Then for each* $\mathtt{r}$ *and any* $\mathtt{x} \geq 1/2$ *, it holds*

$$I\!\!P\Bigl\{ \sup_{\boldsymbol{\upsilon} \in \varUpsilon_{\circ}(\mathtt{r})} \bigl\| \mathcal{Y}(\boldsymbol{\upsilon}) \bigr\| > 6 \nu_0 \mathtt{r} \, z_{\mathbb{H}}(\mathtt{x}) \Bigr\} \leq \mathrm{e}^{-\mathtt{x}},$$

where $z_{\mathbb{H}}(\mathtt{x})$ *is given by the following rule: with* $\mathbb{H} = 4p$

$$z_{\mathbb{H}}(\mathtt{x}) = \begin{cases} \sqrt{\mathbb{H} + 2\mathtt{x}} & \text{if } \mathbb{H} + 2\mathtt{x} \leq \mathtt{g}^2, \\ \mathtt{g}^{-1}\mathtt{x} + \frac{1}{2}\bigl(\mathtt{g}^{-1}\mathbb{H} + \mathtt{g}\bigr) & \text{if } \mathbb{H} + 2\mathtt{x} > \mathtt{g}^2, \end{cases}$$

Outline

Consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$.

PA: Y_i i.i.d. from $P \in (P_{\boldsymbol{\theta}})$ with a log-density $\ell(y, \boldsymbol{\theta})$.

Yields

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\theta}), \\
\widetilde{\boldsymbol{\theta}} &= \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}), \\
\boldsymbol{\theta}^* &= \operatorname*{argmax}_{\boldsymbol{\theta} \in \Theta} \mathbb{E} L(\boldsymbol{\theta}).
\end{aligned}
$$

True: Y_i's are i.i.d. from $P \notin (P_{\boldsymbol{\theta}})$,

$$
D_n^2 = n \mathbb{F}_{\boldsymbol{\theta}^*} .
$$

(for simplicity $p = 1$)

- Smoothness:
  - $\nabla^2 I\!E \ell(Y_1, \boldsymbol{\theta})$ Lipschitz continuous in $\boldsymbol{\theta}$;
  - $I\!E \exp\{\lambda_0 \ell'(Y_1, \boldsymbol{\theta})\} \le \mathtt{C}$
  - $I\!E \exp\{\lambda_0 \ell''(Y_1, \boldsymbol{\theta})\} \le \mathtt{C}$
- Identifiability:

$-\nabla^2 I\!E \ell(\boldsymbol{\theta}) > 0$ and Θ compact;

Then the conditions are fulfilled with $\mathtt{g}^2 \approx n\lambda_0$ and $\mathtt{b}(\mathtt{r}) \ge \mathtt{b}_0 > 0$.

Define $\zeta_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \ell(Y_i, \boldsymbol{\theta}) - I\!E\ell(Y_i, \boldsymbol{\theta})$.

Let

$$\mathbf{v}_0^2 = \operatorname{Var}\bigl\{\nabla\zeta_i(\boldsymbol{\theta}^*)\bigr\}, \qquad V_0^2 = n\mathbf{v}_0^2$$

and

$$\log\exp\bigl\{\lambda n^{-1/2}\mathbf{v}_0^{-1}\nabla\zeta_i(\boldsymbol{\theta}^*)\bigr\} \le \frac{\nu_0^2\lambda^2}{2}, \qquad |\lambda| \le \mathtt{g}_0$$

Then for $|\lambda| \le \mathtt{g}_0 n^{1/2}$

$$\log I\!E\exp\bigl\{\lambda V_0^{-1}\nabla\zeta(\boldsymbol{\theta}^*)\bigr\} = \sum_i \log\exp\bigl\{\lambda n^{-1/2}\mathbf{v}_0^{-1}\nabla\zeta_i(\boldsymbol{\theta}^*)\bigr\} \le \frac{\nu_0^2\lambda^2}{2}.$$

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \sim I\!P$, a sample of independent r.v.s.

Consider PA: $Y_i \sim P_{\Psi_i^\top \boldsymbol{\theta}} \in (P_{\boldsymbol{\upsilon}})$, where

– Ψ_i, given factors in $I\!R^p$,

– $(P_{\boldsymbol{\upsilon}})$, an exponential family with canonical parametrization, $\ell(y, \boldsymbol{\upsilon}) = y\boldsymbol{\upsilon} - d(\boldsymbol{\upsilon})$,

– $\boldsymbol{\theta} \in I\!R^p$, unknown parameter.

MLE:

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}$$

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} I\!E L(\boldsymbol{\theta}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \bigl\{ f_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}$$

with $f_i = I\!E Y_i$.

Checking the conditions

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \bigl\{ Y_i \Psi_i^\top \boldsymbol{\theta} - d(\Psi_i^\top \boldsymbol{\theta}) \bigr\}.$$

Stochastic component is linear in $\boldsymbol{\theta}$

$$\zeta(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) - I\!E L(\boldsymbol{\theta}) = \Bigl(\sum_{i=1}^{n} \varepsilon_i \Psi_i \Bigr)^\top \boldsymbol{\theta}$$

$\nabla^2 \zeta(\boldsymbol{\theta}) \equiv 0$ and (ED_2) automatically;
Fisher information only depends on the model:

$$D_0^2 = \sum_i \Psi_i \Psi_i^\top d''(\Psi_i^\top \boldsymbol{\theta}^*)$$

The vector $\boldsymbol{\xi}$:

$$\boldsymbol{\xi} = D_0^{-1} \nabla \zeta(\boldsymbol{\theta}^*) = D_0^{-1} \sum_{i=1}^{n} \varepsilon_i \Psi_i$$

Sufficient conditions:

– $d''(\Psi_i^\top \boldsymbol{\theta})$ uniformly continuous in $\boldsymbol{\theta}$ over $i = 1, \ldots, n$;

– for some fixed matrices $\mathbf{v}_i^2$ and $\lambda_0 > 0$

$$\mathbb{E} \exp\bigl\{\lambda \mathbf{v}_i^{-1} \varepsilon_i\bigr\} \le \mathtt{C}$$

– the matrix $V_0^2 = \sum_i \mathbf{v}_i^2$ fulfills

$$V_0^2 \le \mathfrak{a}^2 D_0^2$$

The Fisher expansion is simple because the stochastic term is linear in parameter $\boldsymbol{\theta}$. Only smoothness of $d(\boldsymbol{v})$ and exponential moments of Y_i are required.

Median (quantile) regression

Consider a median linear regression

$$Y_i = \Psi_i^\top \boldsymbol{\theta} + \varepsilon_i \,, \qquad \operatorname{med}(\varepsilon_i) = 0.$$

PA: $Y_i - \Psi_i^\top \boldsymbol{\theta} \sim$ i.i.d. Laplace . Yields

$$L(\boldsymbol{\theta}) = -\sum_i |Y_i - \Psi_i^\top \boldsymbol{\theta}| + R$$

MLE = LAD

$$\widetilde{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_i |Y_i - \Psi_i^\top \boldsymbol{\theta}|$$

$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E} L(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_i \mathbb{E} |Y_i - \Psi_i^\top \boldsymbol{\theta}|$$

Checking the conditions

Sufficient conditions:

– the density $f_i(0)$ of $\varepsilon_i = Y_i - \Psi_i^\top \boldsymbol{\theta}^*$ satisfy

$$f_i(0) \geq \mathtt{C} > 0$$

the sample size n satisfies

$$n \geq \mathtt{C} p$$

Observed $Z_i = (X_i, Y_i)$.

Conditional estimating equations (or moment restrictions)

$$\mathbb{E}\big[g(Z,\boldsymbol{\theta}) \,\big|\, X\big] = 0 \quad \text{a.s.} \quad \Leftrightarrow \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0.$$

Here $g(Z,\boldsymbol{\theta})$ is a known function, of Z and $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p$.

Common models that fit into this framework are

1. (non)linear regression models: $g(Z,\boldsymbol{\theta}) = Y - f(X,\boldsymbol{\theta})$;
2. conditional quantile models: $g(Z,\boldsymbol{\theta}) = \mathbb{1}\{Y - f(X,\boldsymbol{\theta}) \le 0\} - \tau$ for a quantile of order τ;
3. linear transformation regression models: $g(Z,\boldsymbol{\theta}) = h(Y,\boldsymbol{\eta}) - X^\top \boldsymbol{\beta}$ and $\boldsymbol{\theta} = (\eta^\top, \boldsymbol{\beta}^\top)^\top$;
4. instrumental variables models;
5. econometric models of optimizing agents, e.g. the consumption model of Hansen and Singleton (1982).

Estimation method

- A classical approach: exploit a finite number of unconditional estimating equations:

$$\mathbb{E}\big[A(X)g(Z,\boldsymbol{\theta}_0)\big] = 0 \quad \text{a.s.}$$

where $A(X)$ is a user-selected matrix function.

- Generalized Method of Moments (GMM) (Hansen, 1982): minimize a weighted quadratic form in the empirical analog of the moment conditions.
- Qin and Lawless (1994) develop an empirical likelihood type estimator.
- Smooth Minimum Distance (SMD) (Lavergne and Patilea, 2010):

$$\mathbb{E}\big[g(Z_1,\boldsymbol{\theta})^\top g(Z_2,\boldsymbol{\theta})\,\omega(X_1 - X_2)\big]\,,$$

where Z_1 and Z_2 are two independent copies of Z , and

$$\omega(x) = \omega_h(x) = K(x/h),$$

where h is a bandwidth and K is a kernel.

Let $\boldsymbol{Z}$ be the observed data. Define

$$M(\boldsymbol{\theta}) = M(\boldsymbol{Z}, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i,j=1}^{n} g_i(\boldsymbol{\theta}) g_j(\boldsymbol{\theta}) w_{ij} \, ,$$

where

- $g_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} g(Z_i, \boldsymbol{\theta})$,
- w_{ij} is the collection of localizing weights: $w_{ij} = N^{-1} K\left(\frac{X_i - X_j}{h}\right)$ and
- N is a normalizing factor which ensures that

$$\sum_j w_{ij} = \frac{1}{N} \sum_j K\left(\frac{X_i - X_j}{h}\right) \approx 1.$$

Simple calculus yields the expectation

$$\mathbb{E} M(\boldsymbol{\theta}) = \sum_{i,j} b_i(\boldsymbol{\theta}) b_j(\boldsymbol{\theta}) w_{ij} + \sum_{i} \mathbb{E} \varepsilon_i^2(\boldsymbol{\theta})\, w_{ii} \,, \tag{9}$$

where $b_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E} g_i(\boldsymbol{\theta})$ and $\varepsilon_i(\boldsymbol{\theta}) \stackrel{\text{def}}{=} g_i(\boldsymbol{\theta}) - \mathbb{E} g_i(\boldsymbol{\theta}) = g_i(\boldsymbol{\theta}) - b_i(\boldsymbol{\theta})$.
Under PA, $\boldsymbol{\theta}^*$ minimizes the first sum in (9):

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i,j} b_i(\boldsymbol{\theta}) b_j(\boldsymbol{\theta}) w_{ij} \,.$$

If the variance $\operatorname{Var} g_i(\boldsymbol{\theta}) = \mathbb{E} \varepsilon_i^2(\boldsymbol{\theta})$ is available, one can consider

$$M^c(\boldsymbol{\theta}) \stackrel{\text{def}}{=} M(\boldsymbol{\theta}) - \sum_{i} \mathbb{E} \varepsilon_i^2(\boldsymbol{\theta})\, w_{ii} \,.$$

Alternatively, one often leaves the cross terms $g_i^2(\boldsymbol{\theta}) w_{ii}$ out in the definition of $M(\boldsymbol{\theta})$

$$M^-(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \sum_{i,j \,:\, i \neq j} g_i(\boldsymbol{\theta}) g_j(\boldsymbol{\theta}) w_{ij} \,.$$

Consider

$$\widetilde{\boldsymbol{\theta}} \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} M^c(\boldsymbol{\theta}) = \sum_{i,j} g_i(\boldsymbol{\theta}) g_j(\boldsymbol{\theta}) w_{ij} - \sum_i I\!E \varepsilon_i^2(\boldsymbol{\theta})\, w_{ii}\,.$$

Define also

$$\boldsymbol{\theta}^* \stackrel{\text{def}}{=} \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} I\!E M^c(\boldsymbol{\theta}) = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Theta} \sum_{i,j} b_i(\boldsymbol{\theta}) b_j(\boldsymbol{\theta}) w_{ij}.$$

Under PA $\boldsymbol{b}(\boldsymbol{\theta}^*) \equiv 0$ and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$.

Aim: accuracy (root-n consistency, efficiency) of $\widetilde{\boldsymbol{\theta}}$.

Problem: the quadratic term $\sum_{i,j} \varepsilon_i(\boldsymbol{\theta}) \varepsilon_j(\boldsymbol{\theta}) w_{ij}$ is not sufficiently regular in $\boldsymbol{\theta}$.

Represent

$$M^c(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta})^\top W \boldsymbol{g}(\boldsymbol{\theta}) - S(\boldsymbol{\theta}) = \|A\boldsymbol{g}(\boldsymbol{\theta})\|^2 - S(\boldsymbol{\theta})$$

where $AA^\top = W$ and $S(\boldsymbol{\theta}) = \sum_i \mathbb{E}\varepsilon_i^2(\boldsymbol{\theta})\, w_{ii}$.

For simplicity suppose that $S(\boldsymbol{\theta})$ is smooth or constant in the vicinity of $\boldsymbol{\theta}^*$.

Define

$$\boldsymbol{g}_0(\boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}(\boldsymbol{\theta}^*).$$

Obviously, with $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{b}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}(\boldsymbol{\theta})$, it holds

$$\sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\{\|A\boldsymbol{g}(\boldsymbol{\theta})\| - \|A\boldsymbol{g}_0(\boldsymbol{\theta})\|\bigr\} \le \sup_{\boldsymbol{\theta}\in\Theta_0(\mathtt{r})} \bigl\|A\bigl\{\boldsymbol{\varepsilon}(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}(\boldsymbol{\theta}^*)\bigr\}\bigr\|.$$

Idea: consider separately $\|A\boldsymbol{g}_0(\boldsymbol{\theta})\|^2$ and $\bigl\|A\bigl\{\boldsymbol{\varepsilon}(\boldsymbol{\theta}) - \boldsymbol{\varepsilon}(\boldsymbol{\theta}^*)\bigr\}\bigr\|$.

Theorem

Suppose that for any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_0)$ *and each* $i = 1, \ldots, n$ *and any unit vector* $\boldsymbol{\gamma} \in I\!\!R^p$

$$\log I\!\!E \exp\Bigl\{ \lambda \boldsymbol{\gamma}^\top \nabla \varepsilon_i(\boldsymbol{\theta}) \Bigr\} \le \nu_0^2 \lambda^2 / 2, \qquad \lambda^2 \le 2 \mathrm{g}^2,$$

Then for each $\mathtt{r}$*, it holds on a random set* $\Omega_1(\mathtt{x})$ *of a dominating probability at least* $1 - \mathrm{e}^{-\mathtt{x}}$

$$\sup_{\boldsymbol{\theta} \in \Theta_0(\mathtt{r})} \| A \boldsymbol{\varepsilon}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) \| \le 6 \nu_0 \, \mathtt{r} \, \mathfrak{z}_A(\mathtt{x}) / \sqrt{N}.$$

where the function $\mathfrak{z}_A(\mathtt{x})$ *is given by*

$$\mathfrak{z}_A(\mathtt{x}) = \mathbb{H}_1 + \sqrt{2\mathtt{x}} + \mathrm{g}^{-1} (\mathrm{g}^{-2} \mathtt{x} + 1) \mathbb{H}_2.$$

Here $\mathbb{H}_1 = 2 \mathbb{H}_1(A)$ *and* $\mathbb{H}_2 = \mathbb{H}_2(A) + 2 \mathfrak{c}_1 p$ *with*

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

he major step in our study is a local linear approximation of $L_0(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\|A\boldsymbol{g}_0(\boldsymbol{\theta})\|^2/2$:

$$L_0(\boldsymbol{\theta}) \stackrel{\mathrm{def}}{=} -\frac{1}{2}\|A\boldsymbol{g}_0(\boldsymbol{\theta})\|^2 = -\frac{1}{2}\big\|A\{\boldsymbol{b}(\boldsymbol{\theta}) + \boldsymbol{\varepsilon}(\boldsymbol{\theta}^*)\}\big\|^2.$$

It is obvious that

$$I\!E L_0(\boldsymbol{\theta}) = -\frac{1}{2}\|A\boldsymbol{b}(\boldsymbol{\theta})\|^2 - \frac{1}{2} I\!E\|A\boldsymbol{\varepsilon}(\boldsymbol{\theta}^*)\|^2.$$

This implies that $\boldsymbol{\theta}^* = \operatorname{argmax}_{\boldsymbol{\theta}} I\!E L_0(\boldsymbol{\theta})$. Further, define

$$D_0^2 = -\nabla^2 I\!E L_0(\boldsymbol{\theta}^*) = -\frac{1}{2}\sum_{i,j}\{\nabla^2 b_i b_j\}(\boldsymbol{\theta}^*) w_{ij} = -\frac{1}{2}\nabla^2\{\boldsymbol{b}^\top W \boldsymbol{b}\}(\boldsymbol{\theta}^*).$$

In the our case when PA is correct, it holds $b_i(\boldsymbol{\theta}^*) \equiv 0$ and

$$D_0^2 = -\frac{1}{2}\sum_{i,j}\nabla b_i(\boldsymbol{\theta}^*)\{\nabla b_j(\boldsymbol{\theta}^*)\}^\top w_{ij} = -\nabla\boldsymbol{b}(\boldsymbol{\theta}^*)^\top W\,\nabla\boldsymbol{b}(\boldsymbol{\theta}^*).$$

Outline

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

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

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

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

Introduce the posterior moments

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

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

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;

Theorem

Suppose the conditions of Theorem 25. Let also $\mathtt{b}(\mathtt{r})$ *from* $(\mathcal{L})$ *satisfies*

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

with $\varrho(\mathtt{r},\mathtt{x})$ *from* (22)*. Then it holds on a random set* $\Omega(\mathtt{x})$ *of probability at least* $1 - 5\mathrm{e}^{-\mathtt{x}}$

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

The bound (10) 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* $\|\boldsymbol{\lambda}\|^2 \le p$

$$\Bigl| \log I\!\!E \Bigl[\exp\bigl\{ \boldsymbol{\lambda}^\top D_0 (\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \bigr\} \bigm| \boldsymbol{Y} \Bigr] - \|\boldsymbol{\lambda}\|^2/2 \Bigr| \le 2\varDelta(\mathtt{r}_0, \mathtt{x}) + 3\mathrm{e}^{-\mathtt{x}},$$

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

$$\begin{aligned} I\!\!P\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \in A \bigm| \boldsymbol{Y} \bigr) &\ge \exp\bigl\{ -2\varDelta(\mathtt{r}_0, \mathtt{x}) - 3\mathrm{e}^{-\mathtt{x}} \bigr\} I\!\!P\bigl(\boldsymbol{\gamma} \in A \bigr) - \mathrm{e}^{-\mathtt{x}}, \\ I\!\!P\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}}) \in A \bigm| \boldsymbol{Y} \bigr) &\le \quad \exp\bigl\{ 2\varDelta(\mathtt{r}_0, \mathtt{x}) + 2\mathrm{e}^{-\mathtt{x}} \bigr\} I\!\!P\bigl(\boldsymbol{\gamma} \in A \bigr) + \mathrm{e}^{-\mathtt{x}}. \end{aligned}$$

- All statements of Theorem 13 require " $\Delta(\mathtt{r}_0, \mathtt{x})$ 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.

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

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

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 $1 - 5\mathrm{e}^{-\mathtt{x}}$

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

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

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

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

Now the result follows from Theorem 13 and

Lemma

Let $\mathbb{P}_0 = \mathcal{N}(0, I_p)$ *and* $\mathbb{P}_1 = \mathcal{N}(\boldsymbol{\beta}, (U^\top U)^{-1})$ *some non-degenerated matrix* U. *If*

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

then $\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) = -\mathbb{E}_0 \log \frac{d\mathbb{P}_1}{d\mathbb{P}_0}$ *fulfills*

$$2\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1) \le \operatorname{tr}(U^\top U - I_p)^2 + (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2 \le \boldsymbol{\epsilon}^2 p + (1 + \boldsymbol{\epsilon})\|\boldsymbol{\beta}\|^2.$$

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

$$\bigl|\mathbb{P}_0(A) - \mathbb{P}_1(A)\bigr| = \bigl|\mathbb{P}\bigl(\boldsymbol{\gamma} \in A\bigr) - \mathbb{P}\bigl(U(\boldsymbol{\gamma} - \boldsymbol{\beta}) \in A\bigr)\bigr| \le \sqrt{\mathcal{K}(\mathbb{P}_0, \mathbb{P}_1)/2}.$$

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 $U^\top U - I_p$. $\|U^\top U - I_p\|_\infty \le \epsilon \le 1/2$ yields $|a_j| \le 1/2$ and

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

This implies by Pinsker's inequality

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

Steps: Local Gaussian approximation of the posterior

Remind

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

and $\boldsymbol{\xi} = D_0^{-1}\nabla L(\boldsymbol{\theta}^*)$. For any nonnegative function f , it holds

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

Similarly,

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

The main benefit of these bounds is that $\mathbb{L}(\boldsymbol{\theta},\boldsymbol{\theta}^*)$ is quadratic in $\boldsymbol{\theta}$.

Theorem

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

$$\mathbb{E}^{\circ}\bigl[f\bigl(D_0(\boldsymbol{\vartheta} - \breve{\boldsymbol{\theta}})\bigr)\,\mathbb{I}_{\mathtt{r}_0}\bigr] \le \exp\bigl\{\Delta^{+}(\mathtt{r}_0, \mathtt{x})\bigr\}\,\mathbb{E} f(\boldsymbol{\gamma}), \qquad (11)$$

where

$$\Delta^{+}(\mathtt{r}_0, \mathtt{x}) = 2\Delta(\mathtt{r}_0, \mathtt{x}) + \nu(\mathtt{r}_0), \qquad (12)$$

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

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

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

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

Upper bound. Proof

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

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

Then

$$m(\boldsymbol{\xi}) + \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = -\|D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\|^2/2 + \log(\det D_0) - p\log(\sqrt{2\pi}) \tag{13}$$

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 (13) for any nonnegative function f that

$$\begin{aligned}
&\int_{\Theta_0(\mathtt{r}_0)} \exp\bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + m(\boldsymbol{\xi})\bigr\}\, f\bigl(D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta} \\
&\qquad \le \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} \int \exp\bigl\{\mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^*) + m(\boldsymbol{\xi})\bigr\}\, f\bigl(D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\bigr)\, d\boldsymbol{\theta} \\
&\qquad = \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})} \int \phi(\boldsymbol{u})\, f(\boldsymbol{u})\, d\boldsymbol{u} = \mathrm{e}^{\Delta(\mathtt{r}_0, \mathtt{x})}\, \mathbb{E} f(\boldsymbol{\gamma}).
\end{aligned} \tag{14}$$

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

$$\begin{aligned} & \int \exp\bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} \, f\bigl(D_0(\boldsymbol{\theta} - \breve{\boldsymbol{\theta}})\bigr) \ \mathbb{I}\bigl\{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \le \mathtt{r}_0\bigr\} d\boldsymbol{\theta} \\ & \quad \ge \exp\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi})\} \int \phi(\boldsymbol{u}) f(\boldsymbol{u}) \ \mathbb{I}\bigl\{\|\boldsymbol{u} + \boldsymbol{\xi}\| \le \mathtt{r}_0\bigr\} d\boldsymbol{u}. \qquad (15) \end{aligned}$$

A special case of (15) with $f(\boldsymbol{u}) \equiv 1$ implies by definition of $\nu(\mathtt{r}_0)$:

$$\int_{\Theta_0(\mathtt{r}_0)} \exp\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\} \, d\boldsymbol{\theta} \ge \exp\bigl\{-\Delta(\mathtt{r}_0, \mathtt{x}) - m(\boldsymbol{\xi}) - \nu(\mathtt{r}_0)\bigr\}. \qquad (16)$$

Now we are prepared to finalize the proof. (14) and (16) imply on $\Omega(\mathtt{r}_0, \mathtt{x})$

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

and (11) follows. As $\|\boldsymbol{\xi}\| \le z(B, \mathtt{x})$ on $\Omega(B, \mathtt{x})$ and $\mathtt{r}_0 \ge z(B, \mathtt{x}) + z(p, \mathtt{x})$,

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

Lemma

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

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

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

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

Obviously $I\!P\bigl\{\boldsymbol{\vartheta} \not\in \Theta_0(\mathtt{r}_0) \bigm| \boldsymbol{Y}\bigr\} \le \rho(\mathtt{r}_0)$. 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

$$\bigl|\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\bigr| \le \mathtt{r}\, \varrho(\mathtt{r}, \mathtt{x}) \tag{17}$$

for $\mathtt{r} = \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|$ and $\varrho(\mathtt{r}, \mathtt{x})$ from (4); cf. the proof of Theorem 25.
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*

$$\mathtt{r}^2\mathtt{b}^2(\mathtt{r}) \ge \mathtt{x} + 2p + 4z^2(B,\mathtt{x}) + 8\mathtt{r}\,\mathtt{b}(\mathtt{r})\varrho(\mathtt{r},\mathtt{x}), \qquad \mathtt{r} \ge \mathtt{r}_0, \tag{18}$$

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

$$\rho(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta\setminus\Theta_0} \exp\{L(\boldsymbol{\theta})\}d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\{L(\boldsymbol{\theta})\}d\boldsymbol{\theta}} \le 2\exp\{-\mathtt{x} + \Delta^{+}(\mathtt{r}_0,\mathtt{x})\} \tag{19}$$

with $\Delta^{+}(\mathtt{r}_0,\mathtt{x})$ *from* (12). *Further, for any unit vector* $\boldsymbol{a} \in I\!\!R^p$

$$\rho_2(\mathtt{r}_0) \stackrel{\text{def}}{=} \frac{\int_{\Theta\setminus\Theta_0} |\boldsymbol{a}^\top D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)|^2 \exp\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\}d\boldsymbol{\theta}}{\int_{\Theta_0} \exp\{L(\boldsymbol{\theta},\boldsymbol{\theta}^*)\}d\boldsymbol{\theta}} \le 2\exp\{-\mathtt{x} + \Delta^{+}(\mathtt{r}_0,\mathtt{x})\}.$$

Suppose that $\mathtt{b}_0 = \mathtt{b}(\mathtt{r}_0)$ is close to one. Condition (18) 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

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

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

$$
-I\!\!E L(\boldsymbol{\theta},\boldsymbol{\theta}^*) \ge \bigl|D_0(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\bigr|^2 \mathtt{b}_k/2
$$

for each $k \ge 0$ and any $\boldsymbol{\theta} \in \Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k)$. The bound (17) implies on $\Omega(\mathtt{x})$

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

for all $k \ge 0$.

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

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

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

Next,

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

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

$$\leq \mathtt{b}_k^{-p/2} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{\mathtt{b}_k} - \frac{1}{4} \mathtt{b}_k \mathtt{r}_k^2 + \frac{p}{2}\Bigr). \tag{20}$$

Here we have used the bound for a standard normal vector $\boldsymbol{\gamma}$ and $\boldsymbol{u} = \mathtt{b}_k^{-1/2} \boldsymbol{\xi} \in I\!R^p$. (16) and (20) imply (19).

Now the bound $\|\boldsymbol{\xi}\| \le z(B, \mathtt{x})$ holding with a dominating probability and (18) imply

$$
\begin{aligned}
&\sum_{k=0}^{\infty} \exp\bigl\{m(\boldsymbol{\xi})\bigr\} \int_{\Theta_0(\mathtt{r}_{k+1}) \setminus \Theta_0(\mathtt{r}_k)} \exp\bigl\{L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)\bigr\} d\boldsymbol{\theta} \\
&\qquad \le \sum_{k=0}^{\infty} \exp\Bigl(\frac{\|\boldsymbol{\xi}\|^2}{\mathtt{b}_k} - \frac{1}{4}\mathtt{b}_k \mathtt{r}_k^2 + \frac{p}{2}\log\bigl(e/\mathtt{b}_k\bigr) + \mathtt{r}_{k+1}\,\varrho(\mathtt{r}_{k+1}, \mathtt{x}) \Bigr) \\
&\qquad \le \sum_{k=0}^{\infty} \exp(-\mathtt{x}/\mathtt{b}_k) \le 2\mathrm{e}^{-\mathtt{x}}
\end{aligned}
$$

and (19) follows in view of $\mathtt{b}\log(e/\mathtt{b}) \le 1$ for $\mathtt{b} \le 1$.

Theorem

Suppose (**??**) *for* $\mathtt{r} = \mathtt{r}_0$ *and* (19). *Then for any nonnegative function* $f(\cdot)$ *on* $\mathbb{R}^p$, *it holds on* $\Omega(\mathtt{x})$

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

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 (14) 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 (15) imply

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

Outline

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.

Penalization

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.

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 $\|G\boldsymbol{\theta}^*_G\|^2$.

“Bias-variance” trade-off:

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

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

Typically V_0^2 measures the local 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 $B_G \stackrel{\mathrm{def}}{=} D_G^{-1} V_0^2 D_G^{-1}$:

$$\mathtt{p}_G \stackrel{\mathrm{def}}{=} \mathrm{tr}\bigl(B_G \bigr).$$

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: $G = \mathrm{diag}\{0, G_1\}$.

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 $g^2/\sigma^2 \gg p_1/p_0$.

For $\beta > 1/2$,

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

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

It holds

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

Define also the index p_e as the largest j satisfying $Lj^\beta \le \sigma$.

$\beta > 1/2$ yields $\mathrm{p}_G \le \mathtt{C}(\beta)\mathrm{p}_e$ 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* $I\!P\bigl(\Omega(\mathtt{x})\bigr) \geq 1 - \mathtt{C}\mathrm{e}^{-\mathtt{x}}$

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

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

with

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

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

Theorem

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

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

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

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*

$$\frac{-2\mathbb{E}L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2} \geq \mathtt{b}(\mathtt{r}), \qquad \forall \boldsymbol{\theta} \in \varTheta_0(\mathtt{r}) = \bigl\{\boldsymbol{\theta} \colon \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\| \leq \mathtt{r}\bigr\}.$$

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

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} 6\nu_0\, z_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\,\omega. \tag{21}$$

Then

$$\mathbb{P}\bigl(\widetilde{\boldsymbol{\theta}} \notin \varTheta_0(\mathtt{r}_0)\bigr) \leq 3\mathrm{e}^{-\mathtt{x}}.$$

$(\mathcal{L}G)$ *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*

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

Theorem

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

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

where

$$\varrho(\mathtt{r}, \mathtt{x}) \stackrel{\mathrm{def}}{=} 6\nu_0\, z_{\mathbb{H}}\bigl(\mathtt{x} + \log(2\mathtt{r}/\mathtt{r}_0)\bigr)\,\omega. \tag{22}$$

Then

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

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

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

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 $\Upsilon_\circ(\mathtt{r}) \stackrel{\mathrm{def}}{=} \{\boldsymbol{\upsilon} \colon \|B^{-1/2}\boldsymbol{\upsilon}\| \le \mathtt{r}\}$

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

Theorem

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

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

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

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

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

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

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

Theorem

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

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

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 $\varOmega(\mathtt{r}, \mathtt{x})$, for each $\boldsymbol{\theta} \in \varTheta_{0,G}(\mathtt{r})$

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

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

Theorem

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

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

where

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

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;

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

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

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

- $p_n^3/n \to 0$ – 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 $\mathbb{F}_G = \mathbb{F}_{\boldsymbol{\theta}_G^*} + n^{-1}G^2$

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

- $\mathtt{p}_G^2/n \to 0$ – 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.

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

Birgé, L. and Massart, P. (1993).
Rates of convergence for minimum contrast estimators.
Probab. Theory Relat. Fields, 97(1-2):113–150.

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.

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.

Portnoy, S. (1984).

Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency.

Ann. Stat., 12:1298–1309.

Portnoy, S. (1988).

Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity.

Ann. Statist., 16(1):356–366.

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 de Geer, S. (1993).

Hellinger-consistency of certain nonparametric maximum likelihood estimators.

Ann. Stat., 21(1):14–44.

van de Geer, S. (2002).

M-estimation using penalties or sieves.
J. Stat. Plann. Inference, 108(1-2):55–69.