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. PA and MLE

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

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

Maximum likelihood estimator (MLE):

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

PA-PW: $I\!\!P \notin (I\!\!P_{\boldsymbol{\theta}})$. 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

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

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

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

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

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 $\operatorname{Cov}(\boldsymbol{Y}) = \Sigma_0$.

Fisher and Wilks for a linear model

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

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 $I\!\!E\boldsymbol{Y}=\boldsymbol{f}$, $\boldsymbol{\varepsilon}\stackrel{\text{def}}{=}\boldsymbol{Y}-\boldsymbol{f}$:

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

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

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

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

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

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

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

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

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:

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

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

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

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

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

The error terms satisfy

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

and

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

I.i.d. case. Dimensional asymptotics $p = p_n \to \infty$

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

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

- $p_n/n \to 0$: 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.

[**?**]: 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 $p^2/n \to \infty$).

[**?**]: 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:

[**?**], [Chen, 1993, 1997], [**?**, **?**]; . . .

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 $\mathtt{r}_0$

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

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

- Local quadratic approximation of the expected log-likelihood:

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

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

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

- 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

$$D^2(\boldsymbol{\theta}) \stackrel{\text{def}}{=} -\nabla^2 \mathbb{E}L(\boldsymbol{\theta}).$$

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

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

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

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

$$\left| -2\mathbb{E}L(\boldsymbol{\theta}) + 2\mathbb{E}L(\boldsymbol{\theta}^*) - \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|^2 \right| \le \delta(\mathtt{r})\mathtt{r}^2,$$

$$\left\| D_0^{-1}\left\{\nabla \mathbb{E}L(\boldsymbol{\theta}) - \nabla \mathbb{E}L(\boldsymbol{\theta}^*)\right\} + D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \right\|$$

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

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

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

$(\boldsymbol{ED_2})$ *There exist a value* $\omega > 0$ *and for each* $\mathtt{r} > 0$, *a constant* $\mathtt{g}(\mathtt{r}) > 0$ *such that* $\zeta(\boldsymbol{\theta}) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - I\!\!E L(\boldsymbol{\theta})$ *satisfies for any* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *:*

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

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

A bound for the norm of a vector stochastic process

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

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

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

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

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

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

Local linear approximation of the gradient

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

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

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

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

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.

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

$$\begin{aligned}\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) &\stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) + \frac{1}{2}\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2 \\ &= L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ), \qquad \boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \varTheta_0(\mathtt{r})\end{aligned}$$

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

$$\bigl|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ)\bigr| = \bigl|(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top D_0 D_0^{-1} \nabla\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)\bigr| \le \|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\| \sup_{\boldsymbol{\theta}' \in \varTheta_0(\mathtt{r})} \bigl|\chi(\boldsymbol{\theta}', \boldsymbol{\theta}^\circ)\bigr| .$$

A bound for the error of the quadratic approximation

$$\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) \stackrel{\text{def}}{=} L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^\circ) - (\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)^\top \nabla L(\boldsymbol{\theta}^\circ) + \frac{1}{2}\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^\circ)\|^2$$

$$= L(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ) - \mathbb{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^\circ), \qquad \boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r})$$

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* $\boldsymbol{\theta}, \boldsymbol{\theta}^\circ \in \Theta_0(\mathtt{r})$

$$\frac{|\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)|}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} \le \diamondsuit(\mathtt{r}, \mathtt{x}), \qquad |\alpha(\boldsymbol{\theta}, \boldsymbol{\theta}^*)| \le \mathtt{r}\,\diamondsuit(\mathtt{r}, \mathtt{x}),$$

$$\frac{|\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})|}{\|D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\|} \le \diamondsuit(\mathtt{r}, \mathtt{x}), \qquad |\alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta})| \le \mathtt{r}\,\diamondsuit(\mathtt{r}, \mathtt{x}).$$

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

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

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 $\mathtt{r} = \mathtt{r}_0$

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

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

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

Together with (1), this yields

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

Square-root Wilks

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 $\boldsymbol{\xi}$.

Large deviations for $\widetilde{\theta}$. Main steps

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

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

- 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

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

$(\mathcal{L})$ *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

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

Then

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

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.

A sharp bound for a norm of a stochastic vector $\boldsymbol{\xi} = D_0^{-1} \nabla L(\boldsymbol{\theta}^*)$

$(\boldsymbol{E D_0})$ *There exist a positive symmetric matrix* V_0^2 *, and constants* $\mathtt{g} > 0$ *,* $\nu_0 \geq 1$ *such that* $\operatorname{Var}\{\nabla\zeta(\boldsymbol{\theta}^*)\} \leq V_0^2$ *and*

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

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

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

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 $\lambda_B = 1$.

A bound for the norm $\|\xi\|$. Cont.

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

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

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

Theorem (SP2012)

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

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

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

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

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

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

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 (2) 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

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

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

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

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

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

Concentration and large deviations. Proof

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

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

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

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

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

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 $\Upsilon_\circ(\mathtt{r}) \stackrel{\text{def}}{=} \{\boldsymbol{\upsilon} : \|H_0(\boldsymbol{\upsilon} - \boldsymbol{\upsilon}^*)\| \le \mathtt{r}\}$.

Theorem

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

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

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

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

Tools. An "upper function" device

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

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

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

Theorem

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

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

Then it holds

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

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 $\mathcal{B}_{\mathtt{r}^*}(\boldsymbol{\upsilon}^*)$ into slices $\mathcal{B}_{\mathtt{r}_k}(\boldsymbol{\upsilon}^*) \setminus \mathcal{B}_{\mathtt{r}_{k-1}}(\boldsymbol{\upsilon}^*)$ and apply Theorem 9 to each slice.
By (3) and Theorem 9 for any $\mathtt{r} > \mathtt{r}_0$

$$\mathbb{P}\biggl(\sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}^*) \setminus \mathcal{B}_{\rho \mathtt{r}}(\boldsymbol{\upsilon}^*)} \bigl\{ \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) - f(\mathtt{r}, \mathtt{r}_0) \bigr\} \geq 0 \biggr)$$

$$\leq \mathbb{P}\biggl(\frac{1}{3\nu_0 \mathtt{r}} \sup_{\boldsymbol{\upsilon} \in \mathcal{B}_{\mathtt{r}}(\boldsymbol{\upsilon}^*)} \mathcal{U}(\boldsymbol{\upsilon}, \boldsymbol{\upsilon}^*) \geq z_{\mathbb{H}}\bigl(\mathtt{x} + \log(\mathtt{r}/\mathtt{r}_0)\bigr) \biggr) \leq \frac{\mathtt{r}_0}{\mathtt{r}} \mathrm{e}^{-\mathtt{x}}. \tag{4}$$

Define $\mathtt{r}_k = \mathtt{r}_0 \rho^{-k}$ for $k = 0, 1, 2, \ldots$ and $k^* \stackrel{\text{def}}{=} \log(\mathtt{r}^*/\mathtt{r}_0) + 1$. By (4)

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

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

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

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 $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ for a fixed point $\boldsymbol{\upsilon}^* \in \varUpsilon$. (w.l.g. $\boldsymbol{\upsilon}^* = 0$).

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

$$\sup_{\boldsymbol{\upsilon} \in \varUpsilon} \log I\!E \exp\Bigl\{ \lambda \boldsymbol{\gamma}^\top \nabla \mathcal{Y}(\boldsymbol{\upsilon}) \boldsymbol{\alpha} \Bigr\} \le \frac{\nu_0^2 \lambda^2}{2}, \qquad \lambda^2 \le 2\mathrm{g}^2. \tag{5}$$

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

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

Condition (5) implies for any $\boldsymbol{\upsilon} \in \varUpsilon_\circ(\mathtt{r})$ with $\|\boldsymbol{\upsilon}\| \le \mathtt{r}$ and $\|\boldsymbol{\gamma}\| = 1$ in view of $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$

$$\log I\!E \exp\Bigl\{ \frac{\lambda}{\mathtt{r}} \boldsymbol{\gamma}^\top \mathcal{Y}(\boldsymbol{\upsilon}) \Bigr\} \le \frac{\nu_0^2 \lambda^2 \|\boldsymbol{\upsilon}\|^2}{2\mathtt{r}^2}, \qquad \lambda^2 \le 2\mathrm{g}^2; \tag{6}$$

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

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

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

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

Suppose that $\boldsymbol{u} \in I\!R^{q}$ and $\boldsymbol{\upsilon} \in \Upsilon$ are such that $\|\boldsymbol{u}\|^2 + \|\boldsymbol{\upsilon}\|^2 \le 2\mathtt{r}^2$. By the Hölder inequality, (6), and (5), it holds for $\|\boldsymbol{\gamma}\| = \|\boldsymbol{\alpha}\| = 1$ and $\boldsymbol{\upsilon} \in \Upsilon_{\circ}(\mathtt{r})$

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

Theorem

Let a random p *-vector process* $\mathcal{Y}(\boldsymbol{\upsilon})$ *for* $\boldsymbol{\upsilon} \in \Upsilon \subseteq I\!R^p$ *fulfill* $\mathcal{Y}(\boldsymbol{\upsilon}^*) = 0$ *,* $I\!E\mathcal{Y}(\boldsymbol{\upsilon}) \equiv 0$ *, and the condition* (5) *be satisfied. Then for each* $\mathtt{r}$ *and any* $\mathtt{x} \geq 1/2$ *, it holds*

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

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

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

Outline

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

Theorem

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

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

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

with

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

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

Main tools

- Expansion of $L(\boldsymbol{\theta}, \boldsymbol{\theta}^*) = L(\boldsymbol{\theta}) - L(\boldsymbol{\theta}^*)$

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

- Taylor of the second order for $I\!E L(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$;
- Local constant approximation of $\nabla\zeta(\boldsymbol{\theta})$ – empirical processes theory for a vector stochastic process; involve the entropy function $z_{\mathbb{H}}(\mathtt{x}) \asymp \sqrt{p + \mathtt{x}}$
- a sharp bound for the squared norm $\|\boldsymbol{\xi}\|^2$; $I\!P\bigl(\|\boldsymbol{\xi}\| \ge z(B, \mathtt{x})\bigr) \le 2\mathrm{e}^{-\mathtt{x}}$ and $z(B, \mathtt{x}) \asymp \sqrt{\operatorname{tr}(B) + \mathtt{x}}$ for the “sandwich” matrix B ;
- “upper function” device for a centered stochastic process (remainder) $\zeta(\boldsymbol{\theta}, \boldsymbol{\theta}^*) - \boldsymbol{\xi}^\top D_0(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\top$ on an unbounded set Θ . Disappears for linear models.

Main steps

- Concentration and large deviations: identify $\mathtt{r}_0$ s.t.

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

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 $\mathtt{r}_0^2 \bigl\{ \delta(\mathtt{r}_0) + \varrho(\mathtt{r}_0, \mathtt{x}) \bigr\}$.

Conditions

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

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

$(E\!\!D_2)$ *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* $\boldsymbol{\theta} \in \Theta_0(\mathtt{r})$ *:*

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

$(E\!\!D_0)$ *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*

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

$(\mathcal{L})$ *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*

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

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

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(Y_i, \boldsymbol{\theta}),$$

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

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

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

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

and

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

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

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

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

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

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

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

$\nabla^2 \zeta(\boldsymbol{\theta}) \equiv 0$ and (ED_2) automatically;

The Fisher information D_0^2 depends on $\mathbb{P}$ only through $\boldsymbol{\theta}^*$:

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

The same for 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 $i = 1, \ldots, n$; here $\ell(y, \boldsymbol{\upsilon}) = y\boldsymbol{\upsilon} - d(\boldsymbol{\upsilon})$;

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

$$\mathbb{E} \exp\bigl\{\lambda_0 \mathbf{v}_i^{-1} \varepsilon_i\bigr\} \le \mathtt{C}$$

– 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 $d(\boldsymbol{\upsilon})$ and exponential moments of Y_i are required.

Consider a median linear regression

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

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

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

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

Sufficient conditions:

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

$$\sup_i \sup_{|u| \le t} \left| \frac{f_i(u)}{f_i(0)} - 1 \right| \le \omega(t), \quad \text{small for } t \text{ small;}$$

the sample size n satisfies

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

Challenges:

- $\nabla L(\boldsymbol{\theta})$ exists but discontinuous;
- $\nabla^2 L(\boldsymbol{\theta})$ is a delta-function;
- $\mathbb{E} L(\boldsymbol{\theta})$ is Lipschitz in $\boldsymbol{\theta}$; we need that $\mathbb{E} L(\boldsymbol{\theta}, \boldsymbol{\theta}^*)$ grows faster than linear.

Observed $Z_i = (X_i, Y_i)$.

Conditional estimating equations (or moment restrictions)

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

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: $g(Z, \boldsymbol{\theta}) = \mathbb{1}\bigl\{Y - f(X, \boldsymbol{\theta}) \le 0\bigr\} - \tau$ 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).

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

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

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):

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

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

$$\mathbb{E} M(\boldsymbol{\theta}) = \sum_{i,j} b_i(\boldsymbol{\theta}) b_j(\boldsymbol{\theta}) w_{ij} + \sum_i \mathbb{E} \varepsilon_i^2(\boldsymbol{\theta})\, w_{ii} \,, \tag{7}$$

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 (7):

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

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 $M(\boldsymbol{\theta})$

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

Estimator

Consider

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

Define also

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

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 $S(\boldsymbol{\theta}) = \sum_i I\!\!E \varepsilon_i^2(\boldsymbol{\theta}) \, w_{ii}$.

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* $\boldsymbol{\gamma} \in I\!\!R^p$

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

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*

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

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

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

A leading term

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

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

It is obvious that

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

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

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

Under PA, it holds $b_i(\boldsymbol{\theta}^*) \equiv 0$ and

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

The result

$$
\begin{aligned}
g_i(\boldsymbol{\theta}) &= b_i(\boldsymbol{\theta}) + \varepsilon_i(\boldsymbol{\theta}), \\
\boldsymbol{\theta}^* &= \operatorname*{arginf}_{\boldsymbol{\theta}} \; \boldsymbol{b}(\boldsymbol{\theta})^\top W \boldsymbol{b}(\boldsymbol{\theta}), \\
\widetilde{\boldsymbol{\theta}} &= \operatorname*{arginf}_{\boldsymbol{\theta}} \{ \boldsymbol{g}(\boldsymbol{\theta})^\top W \boldsymbol{g}(\boldsymbol{\theta}) - S(\boldsymbol{\theta}) \}.
\end{aligned}
$$

Theorem

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

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

the functions $b_i(\boldsymbol{\theta})$ *are twice continuously differentiable uniformly in* i *, and the identifiability condition holds.*
Then $\widetilde{\boldsymbol{\theta}}$ *is root-n consistent and semi parametrically efficient estimator of* $\boldsymbol{\theta}^*$ *.*