European Meeting of Statisticians, Amsterdam, 6-10 July 2015

Semiparametric posterior limits

Bas Kleijn, KdV Institute for Mathematics

Universiteit van Amsterdam

based on joint work with P. Bickel, B. Knapik, M. Chae and Y. Kim

Part I

Regularity, efficiency and semiparametric bias

Example I Regression with symmetric errors

Question

Observe *i.i.d.* $X_1, \ldots, X_n$, $X_i = \theta + e_i$ (or $Y_i = \theta\, X_i + e_i$, *etcetera*) with a symmetrically distributed error. Density for X's is,

$$p_{\theta_0,\eta_0}(x) = \eta_0(x - \theta_0),$$

where $\eta \in H$ is a symmetric Lebesgue density on $\mathbb{R}$. We assume that η is smooth and that the Fisher information for location is non-singular.

Adaptivity Stein (1956), Bickel (1982)

For inference on θ_0 it does not matter whether we know η_0 or not!

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0,\eta_0}\text{-w.}} N(0, I^{-1}_{\theta_0,\eta_0})$$

where I_{θ_0,η_0} is the Fisher information.

Example II Domain boundary estimation

Question

Observe sample $X_1, \ldots, X_n$ *i.i.d.*-P_{θ_0,η_0} with continuous density

$$p_{\theta_0,\eta_0}(x) = \eta_0(x - \theta_0), \quad \eta_0(y) = 0, \text{ if } y < 0.$$

and $\tilde{\lambda}_0 = \eta_0(0) > 0$. Estimate θ_0 with $\eta_0 \in H$, an unknown nuisance.

Model

Define $\mathscr{L} = C_S[0,\infty]$ (cont. $f : [0,\infty] \to \mathbb{R}$ such that $|f| \leq S$)

$$\mathscr{L} \to H : \dot{\ell} \mapsto \eta, \quad \eta(x) = Z_{\dot{\ell}}^{-1}\, e^{-\alpha\, x + \int_0^x \dot{\ell}(t)\, dt}, \quad (Z_{\dot{\ell}} \text{ normalizes})$$

$\eta \in H$ is monotone decreasing, differentiable and log-Lipschitz.

Example III Partial linear regression

Model

Consider *i.i.d.* sample $X_1, \ldots, X_n$, with $X = (Y, U, V) \in \mathbb{R}^3$, related as,

$$Y = \theta_0 U + \eta_0(V) + e$$

where $e \sim N(0,1)$ independent of $(U, V) \sim P$, $\theta_0 \in \mathbb{R}$, $\eta_0 \in H$.

Question

When can we estimate parameter of interest θ_0 in the presence of the unknown nuisance parameter η_0?

Efficiency

There exist estimators $\hat{\theta}_n$ for θ_0 under $P_0 = P_{\theta_0,\eta_0}$ such that,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{P_{\theta_0,\eta_0}\text{-w.}} N(0, \tilde{I}_{\theta_0,\eta_0}^{-1})$$

where $\tilde{I}_{\theta_0,\eta_0}$ is the efficient Fisher information

Likelihood expansions LAN

– local asymptotic normality –

Definition (Le Cam (1960))

There is a $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ with $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$ s.t. for any $(h_n) = O_{P_{\theta_0}}(1)$,

$$\prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h_n}}{p_{\theta_0}}(X_i) = \exp\Big(h_n^T \Delta'_{n,\theta_0} - \tfrac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_{\theta_0}}(1)\Big),$$

where Δ'_{n,θ_0} is given by,

$$\Delta'_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_{\theta_0}\text{-w.}} N(0, I_{\theta_0}),$$

and $I_{\theta_0} = P_{\theta_0}\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T$ is the Fisher information.

Likelihood expansions LAE

– local asymptotic exponentiality –

Definition

There exists a $\tilde{\lambda}_{\theta_0} > 0$ such that for any bounded, stochastic (h_n),

$$\prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1}h_n}}{p_{\theta_0}}(X_i) = \exp\Big(h_n \tilde{\lambda}_{\theta_0} + o_{P_{\theta_0}}(1)\Big)\, 1\{h_n \leq \Delta'_{n,\theta_0}\},$$

where Δ'_{n,θ_0} satisfies,

$$\lim_{n\to\infty} P^n_{\theta_0}(\Delta'_{n,\theta_0} > u) = e^{-\tilde{\lambda}_{\theta_0} u}, \quad \Big(\text{that is } \Delta'_{n,\theta_0} \xrightarrow{P_{\theta_0}\text{-w.}} \mathrm{Exp}^{+}_{0,\tilde{\lambda}_{\theta_0}}\Big)$$

for all $u > 0$. (Ibrigimov and Has'minskii (1981)).

Regular estimation and efficiency I

– definition and convolution theorem –

Definition

An estimator sequence $\widehat{\theta}_n$ for a parameter θ_0 is said to be regular, if for every $h_n = O(1)$, with $\theta_n = \theta_0 + n^{-1/2}h_n$

$$\sqrt{n}(\widehat{\theta}_n - \theta_n) \xrightarrow{P_{\theta_n}\text{-w.}} L_{\theta_0}$$

for some (h_n)-independent limit distribution L_{θ_0}.

Theorem 8.1 *(Hájek, 1970)*
Assume that the model is LAN at θ_0 with non-singular Fisher information I_{θ_0}. Suppose $\widehat{\theta}_n$ is a regular estimator for θ_0 with limit L_{θ_0}. Then there exists a probability kernel M_{θ_0} such that $L_\theta = N(0, I_{\theta_0}^{-1}) * M_{\theta_0}$.

Regular estimation and efficiency II

– asymptotic lineairy and asymptotic bias –

Definition

Given an asymptotic estimation problem with *i.i.d.*-P_0 data and non-singular Fisher information I_0, an influence function Δ_n is,

$$\Delta_n = I_0^{-1}\Delta'_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} I_0^{-1}\dot{\ell}_{\theta_0}(X_i) \xrightarrow{P_0\text{-w.}} N(0, I_0^{-1})$$

Theorem 9.1 *(Fisher, Cramér, Rao, Le Cam, Hájek)*

An estimator $\hat{\theta}_n$ is efficient if and only if it is asymptotically linear:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \Delta_{n,\theta_0} + o_{P_0}(1),$$

for some influence function $\Delta_{n,\theta_0} \xrightarrow{P_{\theta_0}\text{-w.}} N(0, I_{\theta_0}^{-1})$.

Note the asymptotic bias, it equals zero because $P_{\theta_0}\dot{\ell}_{\theta_0} = 0$.

Semiparametric bias

An estimator $\hat{\theta}_n$ for θ_0 is regular but asymptotically biased if,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \tilde{\Delta}_{n,\theta_0,\eta_0} + \mu_{n,\theta_0,\eta_0} + o_{P_0}(1),$$

with $\tilde{\Delta}_{n,\theta_0,\eta_0} \xrightarrow{P_0\text{-w.}} N(0, \tilde{I}^{-1}_{\theta_0\eta_0})$ and $\mu_{n,\theta_0,\eta_0} = O(1)$ or worse. Typically,

$$\left| \mu_{n,\theta_0,\eta_0} \right| \leq \sqrt{n} \sup_{\eta \in D_n} \left| \tilde{I}^{-1}_{\theta_0,\eta_0} P_{\theta_0,\eta} \tilde{\ell}_{\theta_0,\eta_0} \right|$$

where D_n describes some form of localization for $\eta \in H$ around η_0.

Theorem 10.1 *(approximate, see Schick (1986), Klaassen (1987))*
An efficient estimator for θ_0 exists if and only if there exists an estimator $\hat{\Delta}_n$ for the influence function, whose asymptotic bias vanishes at a rate strictly faster than $\sqrt{n}$,

$$P^n_{\theta_n,\eta} \hat{\Delta}_n = o(n^{-1/2}),$$

Part II

Semiparametric posterior limits

Parametric Bernstein-von Mises theorem

Theorem 12.1 (Le Cam (1953), $h = \sqrt{n}(\theta - \theta_0)$)
Let $\mathscr{P} = \{P_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ with thick prior Π_Θ be LAN at θ_0 with non-singular I_{θ_0}. Assume that for every sequence of radii $M_n \to \infty$,

$$\Pi\Big(\|h\| \leq M_n \;\Big|\; X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Then the posterior converges to normality as follows

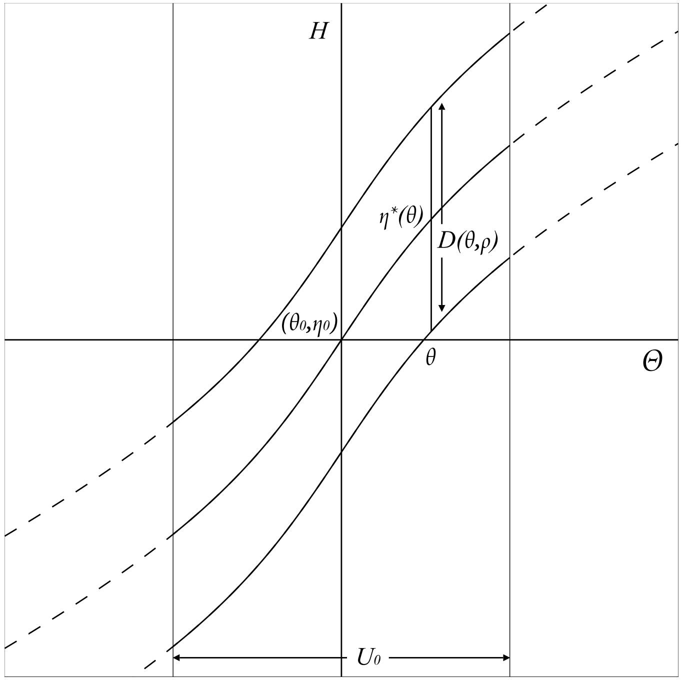
$$\sup_B \Big| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}(B) \Big| \xrightarrow{P_0} 0$$

Another, more familiar form of the assertion,

$$\sup_B \Big| \Pi\Big(\theta \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\hat{\theta}_n, (nI_{\theta_0})^{-1}}(B) \Big| \xrightarrow{P_0} 0$$

for any efficient $\hat{\theta}_n$.

Consistency under $\sqrt{n}$-perturbation

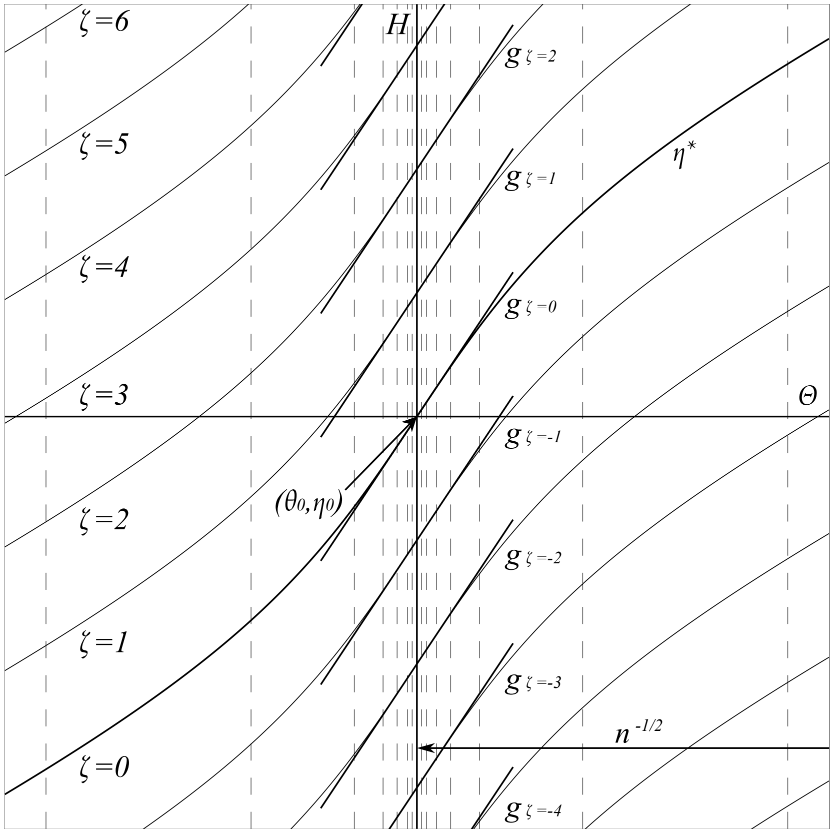


Given $\rho_n \downarrow 0$ we speak of *consistency under $n^{-1/2}$-perturbation at rate ρ_n*, if for all $h_n = O_{P_0}(1)$.

$$\Pi_n\Big(D(\theta, \rho_n) \;\Big|\; \theta = \theta_0 + n^{-1/2} h_n \,;\, X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Integral local asymptotic normality

– Graph/Heuristics –



reparametrize $(\theta, \zeta) \mapsto (\theta, \eta^*(\theta) + \zeta)$

Likelihood expansions ILAN

– Integral local asymptotic normality –

Definition

Given a nuisance prior Π_H, the localized integrated likelihood is,

$$s_n(h) = \int_H \prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h,\eta}}{p_{\theta_0,\eta_0}}(X_i)\, d\Pi_H(\eta),$$

Definition

s_n is said to have the ILAN property, if for every $h_n = O_{P_0}(1)$

$$\log \frac{s_n(h_n)}{s_n(0)} = h_n^T\, \tilde{\Delta}'_{n,\theta_0,\eta_0} - \frac{1}{2} h_n^T \tilde{I}_{\theta_0,\eta_0} h_n + o_{P_0}(1),$$

where the efficient $\tilde{\Delta}'_{n,\theta_0,\eta_0}$ is given by

$$\tilde{\Delta}'_{n,\theta_0,\eta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\infty} \tilde{\ell}_{\theta_0,\eta_0} \xrightarrow{P_{\theta_0,\eta_0}\text{-w.}} N(0, \tilde{I}_{\theta_0,\eta_0})$$

Semiparametric Bernstein-von Mises theorem

Theorem 16.1 (Bickel and BK (2012))

Let $\mathscr{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ with thick prior Π_Θ and nuisance prior Π_H. Assume ILAN at P_{θ_0,η_0} with non-singular $\tilde{I}_{\theta_0,\eta_0}$. Assume that for every sequence of radii $M_n \to \infty$,

$$\Pi\Big(\|h\| \le M_n \;\Big|\; X_1, \ldots, X_n \Big) \xrightarrow{P_0} 1$$

Then the posterior converges marginally to normality as follows

$$\sup_B \Big| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\tilde{\Delta}_{n,\theta_0,\eta_0}, \tilde{I}^{-1}_{\theta_0,\eta_0}}(B) \Big| \xrightarrow{P_0} 0$$

BOTH ILAN and $\sqrt{n}$-consistency are sensitive to semiparametric bias!

Example I Regression with symmetric errors

Theorem 17.1 *(Minwoo Chae, Yongdai Kim and BK (201?))*

Let $X_1, \ldots, X_n$ *be i.i.d.-*P_{θ_0,η_0}*, i.e.* $X_i = \theta_0 + e_i$ *with* e *distributed as a symmetric normal location mixture* η_0 *from* H *of the form,*

$$\eta(x) = \int \phi(x-z)\,dF(z)$$

(where F *is symmetric and* ϕ *denotes the standard normal density). With thick prior* Π_Θ *and nuisance prior* Π_H *that has full weak support, the posterior converges marginally to normality*

$$\sup_B \left| \Pi\Big(h \in B \;\Big|\; X_1, \ldots, X_n \Big) - N_{\tilde{\Delta}_{n,\theta_0,\eta_0}, \tilde{I}^{-1}_{\theta_0,\eta_0}}(B) \right| \xrightarrow{P_0} 0$$

where $\tilde{\ell}_{\theta_0,\eta_0}(X) = \dot{p}_{\theta_0,\eta_0}/p_{\theta_0,\eta_0}(X)$ *and* $\tilde{I}_{\theta_0,\eta_0} = P_0 \tilde{\ell}^2_{\theta_0,\eta_0}$.

Example II Domain boundary estimation

Nuisance prior

Let $S>0$, $W=\{W_s : s\in[0,1]\}$ BM on $[0,1]$, $Z\sim N(0,1)$, indept. of W. Let $\Psi : [0,\infty] \mapsto [0,1]$, $t\mapsto (2/\pi)\arctan(t)$. Define $\dot{\ell}\sim\Pi$ by,

$$\dot{\ell}(t) = S\,\Psi(|Z+W_{\Psi(t)}|).$$

Then $C_S[0,\infty]\subset \mathrm{supp}(\Pi)$.

Theorem 18.1 *(BK and B. Knapik (201?))*

Let $X_1,\ldots,X_n$ be i.i.d.-P_{θ_0,η_0}. Endow $\Theta=\mathbb{R}$ with prior thick at θ_0 and H with prior Π like above. Then,

$$\sup_B \Big| \Pi\Big(h\in B \,\Big|\, X_1,\ldots,X_n\Big) - \mathrm{Exp}^{-}_{\Delta_{n,\theta_0},\tilde{\lambda}_0}(B)\Big| \xrightarrow{P_0} 0$$

where $\Delta_{n,\theta_0}=n(X_{(1)}-\theta_0)$ and $\tilde{\lambda}_0=\eta_0(0)$.

Example III Partial linear regression

Model specification

Observe *i.i.d.*-P_0 sample (U_i, V_i, Y_i) $(i \geq 1)$, $Y = \theta_0 U + \eta_0(V) + e$. where $e \sim N(0,1)$ independent of $(U,V) \sim P$, $PU = 0$, $PU^2 = 1$, $PU^4 < \infty$, $P(U - \mathrm{E}[U|V])^2 > 0$, $P(U - \mathrm{E}[U|V])^4 < \infty$. For given $\alpha > 0$, $M > 0$, define Sobolev ball $H_{\alpha,M} = \{\eta \in C^\alpha[0,1] : \|\eta\|_\alpha < M\}$.

Conjecture 19.1 *(Bickel and BK (2012))*

Let $\alpha > 1/2$ and $M > 0$ be given. Assume that η_0 as well as $v \mapsto \mathrm{E}[U|V=v]$ are in $H_{\alpha,M}$. Let Π_Θ be thick. Choose $k > \alpha - 1/2$ and define $\Pi^k_{\alpha,M}$ to be the distribution of k times integrated Brownian motion started at random, conditioned on $\|\eta\|_\alpha < M$. Then,

$$\sup_A \left| \Pi\Big(h \in A \;\Big|\; X_1, \ldots, X_n\Big) - N_{\tilde{\Delta}_{n,\theta_0,\eta_0}, \tilde{I}^{-1}_{\theta_0,\eta_0}}(A) \right| \xrightarrow{P_0} 0,$$

where $\tilde{\ell}_{\theta_0,\eta_0}(X) = e(U - \mathrm{E}[U|V])$ and $\tilde{I}_{\theta_0,\eta_0} = P(U - \mathrm{E}[U|V])^2$.

The rug from under our feet!

– Why systematic answers are so hard –

Proving ILAN with a product prior $\Pi_\Theta \times \Pi_H$, one has to go through a condition of the form

$$\int_{D_n} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i)\, d\Pi_{H+\delta_n}(\eta) = \bigl(1 + o_{P_0}(1)\bigr) \int_{D_n} \prod_{i=1}^{n} \frac{p_{\theta_0,\eta}}{p_0}(X_i)\, d\Pi_H(\eta)$$

where $\Pi_{H+\delta}(B) := \Pi_H(B+\delta)$ and $\delta_n = \eta^*(\theta_n) - \eta_0$.

For Gaussian prior Π_H approximate δ_n in RKHS and use the Cameron-Martin theorem to obtain the RN-derivative $d\Pi_{H+\delta_n}/d\Pi_H$ explicitly.

For other priors, linearize and estimate δ_n with $\hat{\delta}_n(\theta) = \theta \hat{B}_n$ and choose a nuisance prior Π_H on H and then translate Π_H empirically

$$\Pi_H(\eta \in B | \theta) = \Pi_H(\eta - \hat{\delta}_n \in B)$$

Posterior asymptotic normality

– Analogy/Heuristics –

Parametric posterior

The posterior density $\theta \mapsto d\Pi(\theta | X_1, \ldots, X_n)$

$$\prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta) \Big/ \int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta)$$

with LAN requirement on the likelihood.

Semiparametric analog

The marginal posterior density $\theta \mapsto d\Pi(\theta | X_1, \ldots, X_n)$

$$\int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta) \Big/ \int_\Theta \int_H \prod_{i=1}^{n} p_{\theta,\eta}(X_i)\, d\Pi_H(\eta)\, d\Pi_\Theta(\theta)$$

with integral LAN requirement on Π_H-integrated likelihood.

Marginal convergence at rate $\sqrt{n}$

Theorem 22.1 *(Marginal parametric rate)*

Let Π_Θ and Π_H be given. Assume that there exists a sequence (H_n) of subsets of H, such that the following two conditions hold:

(i) The nuisance posterior concentrates on H_n

$$\Pi\Big(\eta \in H \setminus H_n \,\Big|\, X_1, \ldots, X_n\Big) \xrightarrow{P_0} 0$$

(ii) For every $M_n \to \infty$,

$$\sup_{\eta \in H_n} P_0^n \Pi\Big(n^{1/2}\|\theta - \theta_0\| > M_n \,\Big|\, \eta, X_1, \ldots, X_n\Big) \to 0$$

Then for every $M_n \to \infty$

$$\Pi\Big(n^{1/2}\|\theta - \theta_0\| > M_n \,\Big|\, \eta, X_1, \ldots, X_n\Big) \xrightarrow{P_0} 0$$