Bernstein Von Mises theorem in semi-parametric frameworks

J. Rousseau

CEREMADE, Université Paris-Dauphine

O'Bayes
Joint work with V. Rivoirard

Outline

1. Bernstein Von Mises : general ideas
   - Definition
   - Applications of BVM
   - Conditions
   - Semi-Nonparametric case
   - Our framework
   - Influence function for the CDF
2. An (almost) general result
3. sieve models
   - log- linear sieve models
   - Rates of concentration
   - Condition A2
4. Conclusion

Outline

Bernstein Von Mises : i.i.d parametric

• Observations : for $i = 1, ..., n$ $X_i :\sim f(|\theta)$, i.i.d $\theta \in \Theta$.
A priori : $d\Pi(\theta) = \pi(\theta)d\theta$ = prior distribution
$\longrightarrow$ posterior density

$$\pi(\theta|X^n) = \frac{\pi(\theta)f(X^n|\theta)}{m(X^n)}, \quad X^n = (X_1, ..., X_n)$$

▶ **Bernstein Von Mises :**
When n goes to infinity, the posterior distribution of θ close to a Normal with mean $\hat{\theta}$ and variance $V_{\theta_0}(\hat{\theta})$ under P_{θ_0}.

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, V_{\theta_0}(\hat{\theta})$$

• regular models : $\hat{\theta}$ = MLE, $V_{\theta_0}(\hat{\theta}) = I(\theta_0)^{-1}$ = Inv. Fisher information Matrix

Outline

Applications of BVM

1. Construction of HPD regions

$$C_\alpha^\pi = \{\theta; \pi(\theta|X^n) \geq k_\alpha\}; \quad P^\pi\left[C_\alpha^\pi|X^n\right] = 1 - \alpha$$

Then

$$C_\alpha^\pi \approx \{\theta; (\theta - \hat{\theta})^t J_n(\theta - \hat{\theta}) \leq \mathcal{X}_d^{-1}(1 - \alpha)\}$$

close to the highest likelihood frequentist confidence region.

Applications of BVM

1. Construction of HPD regions

$$C_\alpha^\pi = \{\theta; \pi(\theta|X^n) \geq k_\alpha\}; \quad P^\pi\left[C_\alpha^\pi|X^n\right] = 1 - \alpha$$

   Then

$$C_\alpha^\pi \approx \{\theta; (\theta - \hat{\theta})^t J_n(\theta - \hat{\theta}) \leq \mathcal{X}_d^{-1}(1 - \alpha)\}$$

   close to the highest likelihood frequentist confidence region.

2. α credible regions for θ are asymptotically α-confidence regions

Applications of BVM

1. Construction of HPD regions

$$C_\alpha^\pi = \{\theta; \pi(\theta|X^n) \geq k_\alpha\}; \quad P^\pi\left[C_\alpha^\pi|X^n\right] = 1 - \alpha$$

   Then

$$C_\alpha^\pi \approx \{\theta; (\theta - \hat{\theta})^t J_n(\theta - \hat{\theta}) \leq \mathcal{X}_d^{-1}(1 - \alpha)\}$$

   close to the highest likelihood frequentist confidence region.

2. α credible regions for θ are asymptotically α-confidence regions
3. Approximation of estimators

Outline

Types of conditions required

Theorem

Then

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, I(\theta_0)^{-1})$$

- **Extensions to**

• Non regular models (sometimes)

• Non iid

Types of conditions required

Theorem

1. *If* $\Theta \subset \mathbb{R}^d$

Then

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, I(\theta_0)^{-1})$$

- **Extensions to**
  - Non regular models (sometimes)
  - Non iid

Types of conditions required

Theorem

1. *If* $\Theta \subset \mathbb{R}^d$
2. *If* $f(.|\theta)$ *regular (Positive Fisher, LAN)*

Then

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, I(\theta_0)^{-1})$$

- **Extensions to**

• Non regular models (sometimes)
• Non iid

Types of conditions required

Theorem

1. *If* $\Theta \subset \mathbb{R}^d$
2. *If* $f(.|\theta)$ *regular (Positive Fisher, LAN)*
3. *If* $\forall \epsilon > 0, \exists \delta > 0$ *s.t.*

$$P^n_{\theta_0}\left[\sup_{|\theta-\hat{\theta}_n|>\epsilon}(l_n(\theta) - l_n(\hat{\theta})) \leq -n\delta\right] \to 1$$

Then

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, I(\theta_0)^{-1})$$

▶ **Extensions to**

- Non regular models (sometimes)
- Non iid

Types of conditions required

Theorem

1. *If* $\Theta \subset \mathbb{R}^d$
2. *If* $f(.|\theta)$ *regular (Positive Fisher, LAN)*
3. *If* $\forall \epsilon > 0, \exists \delta > 0$ *s.t.*

$$P^n_{\theta_0}\left[\sup_{|\theta-\hat{\theta}_n|>\epsilon} (l_n(\theta) - l_n(\hat{\theta})) \leq -n\delta\right] \rightarrow 1$$

4. $\pi(\theta_0) > 0$ *and* C^o *in* θ_0

Then

$$\sqrt{n}(\theta - \hat{\theta}) \approx \mathcal{N}(0, I(\theta_0)^{-1})$$

- **Extensions to**

• Non regular models (sometimes)
• Non iid

Why does it work

- **Taylor expansion** of log-likelihood : $l_n(\theta)$ around $\hat{\theta}$

$$\pi(\theta|X^n) \propto e^{l_n(\theta)-l_n(\hat{\theta})+\log(\pi(\theta))-\log(\pi(\hat{\theta}))}$$

$$\propto e^{-\frac{(\theta-\hat{\theta})J_n(\theta-\hat{\theta})}{2}(1+o_P(1))} \quad \text{when } |\theta-\hat{\theta}| = o_P(1)$$

- **Integrate the approximation**

Outline

Semi - parametric case

- **Infinite dimensional :** $\dim(\Theta) = +\infty$
- **Parameter of interest :** $\Psi(\theta) \subset \mathbb{R}^d$
- **Examples :**
  - $\theta = (\psi, \eta)$, $\psi \in \mathbb{R}^d$, $\dim(\eta) = +\infty$: ex. Cox model etc...

$$\Psi(\theta) = \psi$$

Semi - parametric case

- **Infinite dimensional :** $\dim(\Theta) = +\infty$
- **Parameter of interest :** $\Psi(\theta) \subset \mathbb{R}^d$
- **Examples :**
  - $\theta = (\psi, \eta)$, $\psi \in \mathbb{R}^d$, $\dim(\eta) = +\infty$: ex. Cox model etc...

$$\Psi(\theta) = \psi$$

  - $\theta =$ density f, (or regression)

$$\Psi(\theta) = \psi(f), \quad \text{functional}$$

ex : $\psi(f) = F(x) = \int \mathbb{1}_{u \leq x} f(t)dt$, $\psi(f) = \int f^2(u)du$

Existing results

- **Hardly any : some negative**
  - Freedman : infinitely many means + Gaussian prior
    Asymptotic normality but wrong variance – conjugate
- **Recently : Positive results**

Existing results

- **Hardly any : some negative**
  - Freedman : infinitely many means + Gaussian prior
    Asymptotic normality but wrong variance – conjugate
  - Dirichlet Process Prior :
    Asymptotic Normality for CDF – conjugate

- **Recently : Positive results**

Existing results

- **Hardly any : some negative**
  - Freedman : infinitely many means + Gaussian prior
    Asymptotic normality but wrong variance – conjugate
  - Dirichlet Process Prior :
    Asymptotic Normality for CDF – conjugate
  - Kim + Kim and Lee : Cox model + Levy process prior (2005) : $\theta = (\psi, \eta)$
    BVM for ψ and $\int^x d\eta(u)$ – conjugacy
- **Recently : Positive results**

Existing results

- **Hardly any : some negative**
  - Freedman : infinitely many means + Gaussian prior
    Asymptotic normality but wrong variance – conjugate
  - Dirichlet Process Prior :
    Asymptotic Normality for CDF – conjugate
  - Kim + Kim and Lee : Cox model + Levy process prior (2005) : $\theta = (\psi, \eta)$
    BVM for ψ and $\int^x d\eta(u)$ – conjugacy
- **Recently : Positive results**
  - Shen (02, JASA) : General conditions BUT
    - a mess
    - mistakes + condns very hard to verify

Existing results

- **Hardly any : some negative**
  - Freedman : infinitely many means + Gaussian prior
    Asymptotic normality but wrong variance – conjugate
  - Dirichlet Process Prior :
    Asymptotic Normality for CDF – conjugate
  - Kim + Kim and Lee : Cox model + Levy process prior (2005) : $\theta = (\psi, \eta)$
    BVM for ψ and $\int^x d\eta(u)$ – conjugacy
- **Recently : Positive results**
  - Shen (02, JASA) : General conditions BUT
    - a mess
    - mistakes + condns very hard to verify
  - Castillo (08) : $\theta = (\psi, \eta)$
    - Gives general conditions to prove BVM for ψ
    - Clarifies Shen
    - Applications in Cox model, White noise model

Outline

Our framework

- **Model :** $X^n = (X_1, ..., X_n)$ i.i.d f unknown on $\mathbb{R}$.

  π : prior on f
- **Parameter of interest :** Continuous Linear functional of f

$$\Psi(f) = \int_{\mathbb{R}} \psi(u) f(u) du$$

ex : $\Psi(f) = F(x), \quad \psi(u) = \mathbb{I}_{u \leq x}$

- **Aim :** Asymptotic posterior distribution of $\Psi(f)$:
  - Normality ?

Our framework

- **Model :** $X^n = (X_1, ..., X_n)$ i.i.d f unknown on $\mathbb{R}$.
  π : prior on f
- **Parameter of interest :** Continuous Linear functional of f

$$\Psi(f) = \int_{\mathbb{R}} \psi(u) f(u) du$$

ex : $\Psi(f) = F(x), \quad \psi(u) = \mathbb{I}_{u \leq x}$

- **Aim :** Asymptotic posterior distribution of $\Psi(f)$:
  - Normality ?
  - Relation with frequentist distribution of

$$\Psi(\mathbb{P}_n) = \frac{\sum_{i=1}^n \psi(X_i)}{n}$$

Do we have :

$$\sqrt{n}(\Psi(f) - \Psi(\mathbb{P}_n)) \approx \mathcal{N}(0, V(\Psi(X))??$$

ex :

$$\sqrt{n}(F(x) - F_n(x)) \approx \mathcal{N}(0, F_0(x)(1 - F_0(x)), \quad ??$$

Outline

Influence function

Control of

$$E^{\pi}\left[e^{t\sqrt{n}(\Psi(f)-\Psi(\mathbb{P}_n))}|X^n\right]$$

- **Influence function $\tilde{\psi}$**

If $f_h \approx f_0(1+hg)$, $u \approx 0$

$$\psi(f_h) = \psi(f_0) + h\int f_0 g\tilde{\psi} + o(h)$$

ex : CDF

$$\tilde{\psi}(u) = \mathbb{1}_{u\le x} - F_0(u)$$

- **idea**

$$l_n(f) - l_n(f_0) + t\sqrt{n}(\Psi(f) - \Psi(\mathbb{P}_n)) = \frac{t^2 V_0}{2}$$
$$+ l_n(f \times e^{-t\bar{\psi}/\sqrt{n}}) - l_n(f_0) + o_P(1)$$

An (almost) general result

$$E^{\pi}\left[e^{t\sqrt{n}(\psi(f)-\psi(\mathbb{P}_n)}|X^n\right] \xrightarrow{?} e^{\frac{t^2}{2}V_0}, \quad F_0 = \text{true}$$

- **Conditions**
  - **[A1]** Let $A_{\epsilon_n} = \{f, d(f_0, f) \leq \epsilon_n\}$, $\epsilon_n = o(1)$ with $d(f, f') = \|\log f - \log f'\|_f$

$$P^{\pi}\left[A_{\epsilon_n}|X^n\right] \to 1, \quad \text{and}$$

Then

$$P^{\pi}\left[\sqrt{n}(\psi(f) - \psi(\mathbb{P}_n)) \leq z|X^n\right] \longrightarrow \Phi\left(\frac{z}{\sqrt{V_0}}\right)$$

An (almost) general result

$$E^{\pi}\left[e^{t\sqrt{n}(\psi(f)-\psi(\mathbb{P}_n)}|X^n\right] \xrightarrow{?} e^{\frac{t^2}{2}V_0}, \quad F_0 = \text{true}$$

- **Conditions**
  - **[A1]** Let $A_{\epsilon_n} = \{f, d(f_0, f) \leq \epsilon_n\}$, $\epsilon_n = o(1)$ with $d(f, f') = \| \log f - \log f' \|_f$

$$P^{\pi}[A_{\epsilon_n}|X^n] \to 1, \quad \text{and}$$

  - **[A2]**

$$\frac{\int_{A_{\epsilon_n}} f \times e^{-t\tilde{\psi}/\sqrt{n}}(X^n)d\pi(f)}{\int_{A_{\epsilon_n}} f(X^n)d\pi(f)} = 1 + o_{p_0}(1),$$

Then

$$P^{\pi}\left[\sqrt{n}(\psi(f) - \psi(\mathbb{P}_n)) \leq z|X^n\right] \longrightarrow \Phi\left(\frac{z}{\sqrt{V_0}}\right)$$

Comments

- **[A1]** Usual type of condition. Posterior concentration rates

Comments

- **[A1]** Usual type of condition. Posterior concentration rates
- **[A2]** Means that we can consider a *change of parameters*

$$f' = f \times e^{-\frac{t\tilde{\psi}}{\sqrt{n}}}, \quad s.t. \quad d\pi'(f') = d\pi(f)(1 + o(1))$$

In parametric cases : $\theta' = \theta + tu/\sqrt{n}$

$$\pi(\theta') = \pi(\theta)(1 + o(1)), \quad \text{if } \pi \quad \text{is } C^o$$

In nonparametric : "holes" in π.

Why it could not work

- **Sieve models**

$$\mathcal{F} = \cup_k \mathcal{F}_k, \quad \mathcal{F}_k \subset \mathcal{F}_{k+1}, \quad \mathcal{F}_k = \{f_{\theta_k}, \theta_k \in \Theta_k\}$$

$$d\pi(f) = p(k) d\pi_k(\theta), \quad f = f_\theta, \quad \theta \in \Theta_k / \Theta_{k-1}$$

- **If** $f_0 \in \mathcal{F}_{k^*}$ for some $k^* \in \mathbb{N}$

Then

$$P^\pi[k^* | X^n] = 1 + o_P(1)$$

and

$$F(x) \approx \mathcal{N}(F_{\hat{\theta}_{k^*}}(x), V_{k^*}(\theta_0)/n)$$

- **No BVM wrt** $F_n(x)$
- **Sieve models : between parametric and non parametric** - depends on f_0

Outline

Log- linear sieve models on $[0, 1]$

$$\mathcal{F} = \{f; \log f \in L^2([0,1])\}, \quad f_0 \in \mathcal{F}$$

- **Representation in sieves**

$$f_\theta(u) = \exp\{\sum_{j=0}^{k} \theta_j \phi_j(u) - c(\theta)\}, \quad (\phi_j)_j^\infty \equiv \text{BON}, \quad \phi_0 = 1$$

ex : Fourier, wavelets, log-spline.

- **Prior** $d\pi(f_\theta) = p(k) d\pi_k(\theta)$

• random sieves : $\mathrm{supp(p)} = \mathbb{N}$ (Poisson, Negative Bino, Geometric)

• deterministic : $p(k) = \delta_{k_n}(k)$, $\lim_n k_n = +\infty$

Outline

Rates of concentration : cond **[A1]**

- **Condition on f_0**

$$\|\log f_0\|_\infty < +\infty, \quad \exists \gamma > 1/2 \quad s.t. \quad \sum_j j^{2\gamma}\theta_{0j}^2 < \infty \quad (S_\gamma, \mathcal{B}^\gamma_{p,q}, p \geq 2)$$

- **Conditions on the prior**
  - On $\mathcal{F}_k : \forall j \leq k,\ 0 < \tau_j = \tau_0 j^{-2\beta}$

$$\frac{\theta_j}{\sqrt{\tau_j}} \overset{iid}{\sim} g, \quad g(x) \leq Ce^{-|x|^a}, \quad a > 0$$

Rates of concentration : cond **[A1]**

- **Condition on f_0**

$$\|\log f_0\|_\infty < +\infty, \quad \exists \gamma > 1/2 \quad s.t. \quad \sum_j j^{2\gamma}\theta_{0j}^2 < \infty \quad (S_\gamma, \mathcal{B}_{p,q}^\gamma, p \geq 2)$$

- **Conditions on the prior**
  - On $\mathcal{F}_k : \forall j \leq k$, $0 < \tau_j = \tau_0 j^{-2\beta}$

$$\frac{\theta_j}{\sqrt{\tau_j}} \overset{iid}{\sim} g, \quad g(x) \leq Ce^{-|x|^a}, \quad a > 0$$

  - $\beta > 1/2$ and ϕ_j uniformly bounded.

Rates of concentration : cond **[A1]**

- **Condition on f_0**

$$\|\log f_0\|_\infty < +\infty, \quad \exists \gamma > 1/2 \quad s.t. \quad \sum_j j^{2\gamma}\theta_{0j}^2 < \infty \quad (S_\gamma, \mathcal{B}_{p,q}^\gamma, p \geq 2)$$

- **Conditions on the prior**
  - On $\mathcal{F}_k$: $\forall j \leq k$, $0 < \tau_j = \tau_0 j^{-2\beta}$

$$\frac{\theta_j}{\sqrt{\tau_j}} \overset{iid}{\sim} g, \quad g(x) \leq Ce^{-|x|^a}, \quad a > 0$$

  - $\beta > 1/2$ and ϕ_j uniformly bounded.
  - Prior on k
    - either : $e^{-c_1 k \log k} \leq p(k) \leq e^{-c_2 k}$
    - Or $p(k) = \delta_{k_n}(k)$, $k_n = k_0 n^{1/(2\beta+1)} \log n$

Results

Under the above types of prior : (same types for wavelets)

- If k random

$$P^{\pi}\left[f_{\theta};\|\theta-\theta_0\|_2 \leq Mn^{-\gamma/(2\gamma+1)}\log n^q|X^n\right] \rightarrow 1$$

and **[A1]** is satisfied + adaptive rate of convergence on the

$$\cup_{\gamma>1/2}S_{\gamma}$$

Generalisation of C. Scricciolo - different approach to adaptation

Results

Under the above types of prior : (same types for wavelets)

- If k random

$$P^{\pi}\left[f_\theta;\, \|\theta-\theta_0\|_2 \leq M n^{-\gamma/(2\gamma+1)} \log n^q | X^n\right] \to 1$$

  and **[A1]** is satisfied + adaptive rate of convergence on the

$$\cup_{\gamma>1/2} S_\gamma$$

- If $k = k_n$

$$P^{\pi}\left[f_\theta;\, \|\theta-\theta_0\|_2 \leq M n^{-(\beta\wedge\gamma/(2\beta+1)} \log n^q | X^n\right] \to 1$$

  and **[A1]** is satisfied but NON adaptive rate of convergence on the

$$\cup_{\gamma>1/2} S_\gamma$$

Generalisation of C. Scricciolo - different approach to adaptation

Outline

Condition [A2]

$$\frac{\int_{A_{\epsilon_n}} f \times e^{-t\tilde{\psi}/\sqrt{n}}(X^n)d\pi(f_h)}{\int_{A_{\epsilon_n}} f(X^n)d\pi(f)} = 1 + o_{p_0}(1),$$

- If $a \geq 1$ $(g(x) \leq Ce^{-|x|})$ and if $k = k_n$

  BVM is valid

Condition [A2]

$$\frac{\int_{A_{\epsilon n}} f \times e^{-t\tilde{\psi}/\sqrt{n}}(X^n)d\pi(f_h)}{\int_{A_{\epsilon n}} f(X^n)d\pi(f)} = 1 + o_{p_0}(1),$$

- If $a \geq 1$ ($g(x) \leq Ce^{-|x|}$) and if $k = k_n$
  BVM is valid
- If $a \geq 1$ and k random then

$$P^{\pi}\left[\sqrt{n}(F(x) - F_n(x)) \leq z|X^n\right] = \sum_k p(k|X^n)\Phi\left(\frac{z + \sqrt{n}\mu_{n,k}}{\sqrt{V_0}}\right)$$

where $\mu_{n,k} = 0(\sum_{k+1}^{\infty} \theta_{0j}\psi_j)$, if $\mathbb{I}_{u \leq x} = \sum_j \psi_j \phi_j$

Explanations on $\mu_{n,k}$

- **change of parameter**

$$f'_\theta = f_\theta e^{-t\bar{\psi}/\sqrt{n}} = e^{\sum_{j=0}^{k}(\theta_j - t\bar{\psi}_j/\sqrt{n})\phi_j(u) - t\sum_{j\geq k+1}\bar{\psi}_j\phi_j/\sqrt{n}}$$

- In $\mathcal{F}_k$ possible change of parameters : $\theta'_j = \theta_j - t\bar{\psi}_j/\sqrt{n}, j \leq k$
- But $t\sum_{j\geq k+1}\bar{\psi}_j\phi_j/\sqrt{n}$ Might be "too big" if k not large enough.

$$\mu_{n,k} = F_0[(\tilde{\psi} - M_{f_0,k}\tilde{\psi})\sum_{j\geq k+1}\theta_{0j}\phi_j]$$

An example where it does not work

$$f_0 \propto \exp\left(\sum_{j=k_0}^{\infty} \theta_{0j}\phi_j\right) \quad \theta_{0,2j} = 0 \quad \theta_{0,2j-1} = \frac{\sin(2\pi jx)}{[j^{\gamma+1/2}\sqrt{\log j}\log\log j]}$$

Implies that for

$$c_1\sqrt{n}k^{-\gamma-1/2} \leq |\mu_{n,k}| \leq \sqrt{n}c_2k^{-\gamma-1/2}$$

Moreover

$$P^{\pi}[k \leq k_n|X^n] = 1 + o(1)$$

with

$$k_n = k_0\left(\frac{n}{(\log n \log\log n)^2}\right)^{1/(2\gamma+1)} \Rightarrow \mu_{n,k} = 0(\log n)$$

conclusion

- BVM in semi-parametric (non parametric) : not obvious...

conclusion

- BVM in semi-parametric (non parametric) : not obvious...
- No easy conditions : because need of change of parameters

conclusion

- BVM in semi-parametric (non parametric) : not obvious...
- No easy conditions : because need of change of parameters
- More work to extend to non log-linear sieve priors ex : mixtures ? ? ?

conclusion

- BVM in semi-parametric (non parametric) : not obvious...
- No easy conditions : because need of change of parameters
- More work to extend to non log-linear sieve priors ex : mixtures ? ? ?
- But very interesting