# Technics for studying asymptotic properties of the posterior distribution in high dimensional models

J. Rousseau

CEREMADE, Université Paris-Dauphine & ENSAE - CREST

JDS, Montpellier

# Outline

1. Bayesian statistics
   - General setup
2. Bayesian nonparametrics
3. On the consistency and posterior concentration rates
   - Definitions
4. Empirical Bayes
   - Driving example
   - Change of measure
5. Application to DP mixtures of Gaussians
6. On frequentist properties of credible regions
7. Semi - parametric : BvM
   - Semi-parametric Bayesian methods
   - BvM in the parametric case
   - Applications of BVM
   - Conditions
   - A General BVM theorem
   - A direct approach
   - General theorem : a nasty condition

# Outline

# Bayesian statistics

- **Sampling model and prior models**
- $X^n|\theta \sim P_\theta$ on $\mathcal{X}_n$ with $\theta \in \Theta$
- $\theta$ : unknown $\rightarrow$ random variable . $\Pi$ = prior proba on $(\Theta, \mathcal{A})$
- **joint, marginal and posterior distributions**
- Joint $(X^n, \theta) \sim P_\theta \times \Pi$
- Posterior : $\Pi(d\theta|X^n)$ If dominated model $f_\theta = dP_\theta/d\mu$

$$\Pi(d\theta|X^n) = \frac{f_\theta(X^n)\Pi(d\theta)}{m(X^n)}, \quad m(X^n) = \int_\Theta f_\theta(X^n)\Pi(d\theta)$$

- Marginal of $X^n$ : $m(X^n)$

# Examples

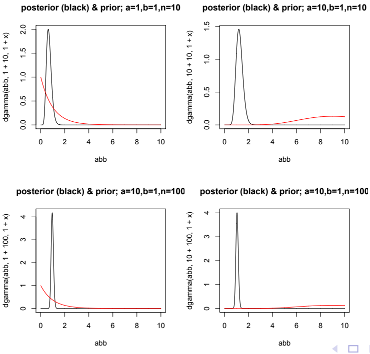
- **Parametric**

Poisson model : $X^n = (X_1, \cdots, X_n)$, $X_i \sim \mathcal{P}(\theta)$

Prior on $\theta > 0$ $\Gamma(a, b)$

- Posterior

$$\Pi(\theta|X^n) \equiv \Gamma(a + n, b + n\bar{X}_n), \quad \bar{X}_n = \sum_i X_i/n$$



# What do we observe ?

- Posterior concentration : posterior shrinks towards $\theta_0 = 1$

# What do we observe ?

- Posterior concentration : posterior shrinks towards $\theta_0 = 1$
- Prior becomes less and less influential as $n \uparrow$.

# What do we observe ?

- Posterior concentration : posterior shrinks towards $\theta_0 = 1$
- Prior becomes less and less influential as $n \uparrow$.
- Asymptotic normality of the posterior : BvM

# What do we observe ?

- Posterior concentration : posterior shrinks towards $\theta_0 = 1$
- Prior becomes less and less influential as $n \uparrow$.
- Asymptotic normality of the posterior : BvM
- General features in regular parametric models

# What do we observe ?

- Posterior concentration : posterior shrinks towards $\theta_0 = 1$
- Prior becomes less and less influential as $n \uparrow$.
- Asymptotic normality of the posterior : BvM
- General features in regular parametric models
- How can we extend these results in large dimensional models ?

# First : posterior distribution = more than point estimation

- **What can we do with the posterior distribution ?**
  - Point estimators : Loss function : $\ell : \Theta \times \mathcal{D} \rightarrow \mathbb{R}^+$
    Bayes estimator

$$\delta^\pi(X^n) = \text{argmin}_\delta E^\pi \left[\ell(\theta, \delta) | X^n\right]$$

e.g. $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ then $\delta^\pi(X^n) = E^\pi\left(\theta | X^n\right)$.

# First : posterior distribution = more than point estimation

- **What can we do with the posterior distribution ?**
  - Point estimators : Loss function : $\ell : \Theta \times \mathcal{D} \to \mathbb{R}^+$
    Bayes estimator

$$\delta^\pi(X^n) = \text{argmin}_\delta E^\pi\left[\ell(\theta, \delta)|X^n\right]$$

    e.g. $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ then $\delta^\pi(X^n) = E^\pi(\theta|X^n)$.
  - Credible regions : measure of uncertainty

$$C_\alpha : \Pi\left(\theta \in C_\alpha|X^n\right) \geq 1 - \alpha$$

# First : posterior distribution = more than point estimation

- **What can we do with the posterior distribution ?**
  - Point estimators : Loss function : $\ell : \Theta \times \mathcal{D} \to \mathbb{R}^+$
    Bayes estimator

    $$\delta^\pi(X^n) = \text{argmin}_\delta E^\pi\left[\ell(\theta, \delta)|X^n\right]$$

    e.g. $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ then $\delta^\pi(X^n) = E^\pi\left(\theta|X^n\right)$.
  - Credible regions : measure of uncertainty

    $$C_\alpha : \Pi\left(\theta \in C_\alpha|X^n\right) \geq 1 - \alpha$$

  - Risk estimation

    $$\hat{R} = E^\pi\left(\ell(\theta, \hat{\delta}_n)|X^n\right)$$

# First : posterior distribution = more than point estimation

- **What can we do with the posterior distribution ?**
  - Point estimators : Loss function : $\ell : \Theta \times \mathcal{D} \to \mathbb{R}^+$
    Bayes estimator

$$\delta^\pi(X^n) = \text{argmin}_\delta E^\pi \left[\ell(\theta, \delta)|X^n\right]$$

    e.g. $\ell(\theta, \theta') = \|\theta - \theta'\|_2^2$ then $\delta^\pi(X^n) = E^\pi(\theta|X^n)$.
  - Credible regions : measure of uncertainty

$$C_\alpha : \Pi\left(\theta \in C_\alpha|X^n\right) \geq 1 - \alpha$$

  - Risk estimation

$$\hat{R} = E^\pi\left(\ell(\theta, \hat{\delta}_n)|X^n\right)$$

  - testing : e.g.

$$\Pi(\Theta_0|X^n) > \Pi(\Theta_1|X^n) \quad \Leftrightarrow \quad \text{accept} \quad \Theta_0$$

# Questions

- What can we say about

$$E_{\theta_0}\left[\ell(\theta_0, \hat{\delta}^{\pi}(X^n))\right]?$$

- What can we say about

$$P_{\theta_0}\left[\theta_0 \in C_{\alpha}\right]?$$

- What can we say about

$$P_{\theta}[\Pi(\Theta_0|X^n) > \Pi(\Theta_1|X^n)]?$$

# Questions

- What can we say about

$$E_{\theta_0}\left[\ell(\theta_0, \hat{\delta}^{\pi}(X^n))\right]?$$

Standard using posterior concentration rates

- What can we say about

$$P_{\theta_0}\left[\theta_0 \in C_\alpha\right]?$$

Difficult

# Bayesian nonparametrics

- **Setup** $\Theta$ is infinite dimensional.
- **Examples**

• Regression function : $Y_i = f(X_i) + \epsilon_i$, $f : \mathbb{R}^d \to \mathbb{R}$

$$\Theta = L_2$$

• Density estimaton $Y_i \overset{iid}{\sim} f$

$$\Theta = \mathcal{F} = \{f : \mathbb{R}^d \to \mathbb{R}^+, \int f = 1\}$$

• classification , spectral density , intensity , conditional density, etc . . .

# Examples of priors : Gaussian process priors

- **Gaussian process priors** $(\Theta, \|.\|)$ Banach space (e.g. $L_2$) $\theta = f$

$$f \sim GP(0, K), \quad \Rightarrow (f(r_1), \cdots, f(r_q)) \sim \mathcal{N}(0, (K(r_i, r_j))_{i,j \leq q})$$

$K$ : drives the smoothness of $f$.

- $K(r, s) = \min(s, t)$ : Brownion – Non statio., non smooth

- **Serie representation** [Karhunen Loeve expansion] : Hilbert Space

$$f = \sum_{i=1}^{\infty} \theta_i \phi_i, \quad (\phi_i)_i = \text{BON } \mathbb{H} \quad \theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2), \quad \tau_i \downarrow 0$$

• good for curves in $\mathbb{R}$ – not so good for densities , etc.

# Examples of priors : Gaussian process priors

► **Gaussian process priors** $(\Theta, \|.\|)$ Banach space (e.g. $L_2$)
$\theta = f$

$$f \sim GP(0, K), \quad \Rightarrow (f(r_1), \cdots, f(r_q)) \sim \mathcal{N}(0, (K(r_i, r_j))_{i,j \leq q})$$

$K$ : drives the smoothness of $f$.

- $K(r, s) = \min(s, t)$ : Brownion – Non statio., non smooth
- $K(r, s) = e^{-a(r-s)^2}$ : exponential kernel – statio. , smooth

► **Serie representation** [Karhunen Loeve expansion] : Hilbert Space

$$f = \sum_{i=1}^{\infty} \theta_i \phi_i, \quad (\phi_i)_i = \text{BON } \mathbb{H} \quad \theta_i \overset{ind}{\sim} \mathcal{N}(0, \tau_i^2), \quad \tau_i \downarrow 0$$

• good for curves in $\mathbb{R}$ – not so good for densities , etc.

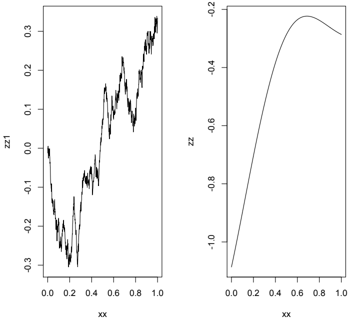



FIG.: Gaussian processes : left : Brownian motion, right : exponential

# Other priors on curves in $\mathbb{R}$ : hierarchical modelling

▶ **Splines, basis expansions**

$$f = \sum_{i=1}^{K} \theta_i \phi_i, \quad (\phi_i)_i = \text{Base } \mathbb{H} \quad \theta_i/\tau_i \overset{iid}{\sim} g(.)$$

• Choice of $K$ , of $\tau_i$ of $g$ ?

- <u>$K$ random</u> : $K \sim \Pi_K$ ; then $\tau_i = \tau$ is enough – $g$ flexible

$\longrightarrow$ more flexible - adaptation to the smoothness

# Other priors on curves in $\mathbb{R}$ : hierarchical modelling

- **Splines, basis expansions**

$$f = \sum_{i=1}^{K} \theta_i \phi_i, \quad (\phi_i)_i = \text{Base } \mathbb{H} \quad \theta_i / \tau_i \overset{iid}{\sim} g(.)$$

• Choice of $K$ , of $\tau_i$ of $g$ ?

- $\underline{K \text{ random}}$ : $K \sim \Pi_K$ ; then $\tau_i = \tau$ is enough – $g$ flexible
- $\underline{\tau_i = \tau(1+i)^{-\alpha-1/2}}$, $K = +\infty$, $g = \mathcal{N}$
  either $\tau \sim \pi_\tau$ or $\alpha \sim \pi_\alpha$ or EB

$\longrightarrow$ more flexible - adaptation to the smoothness

# Nonparametric mixture models

- **Density modelling**

$$f_{P,\sigma} = K_\sigma P(x) = \int_\Theta g_{\theta,\sigma}(x) dP(\theta), \quad P = \text{proba}$$

e.g.

$$g_{\theta,\sigma} = \mathcal{N}(.|\theta,\sigma), \quad \text{or } \mathcal{N}(.|\mu,\tau^2), \quad \theta = (\mu,\tau^2)$$

- **Prior** $P \sim \Pi_P$ and $\sigma \sim \pi_\sigma$
- **Examples of** $\Pi_P$

• finite mixtures :

$$P = \sum_{j=1}^{K} p_j \delta_{(\theta_j)}, \quad K \sim \Pi_K, \ (p_1, \cdots, p_k)|K = k \sim \pi_{p|k}, \quad \theta_j \overset{iid}{\sim} \pi_\theta$$

• Dirichlet Process and co.

# Dirichlet Process : $P \sim DP(M, G)$

- **Sethuraman representation**

$$P = \sum_{i=1}^{\infty} p_j \delta_{(\theta_j)}, \quad \theta_j \overset{iid}{\sim} G,$$

$$p_j = V_j \prod_{i<j} (1 - V_i), \quad V_j \overset{iid}{\sim} Beta(1, M) : \text{ stick breaking}$$

- **Partition property** $\forall (B_1, \cdots, B_k)$ partition

$$(P(B_1), \cdots, P(B_k)) \sim \mathcal{D}(MG(B_1), \cdots, MG(B_k))$$

- **Nice clustering properties** Chinese restaurant process.

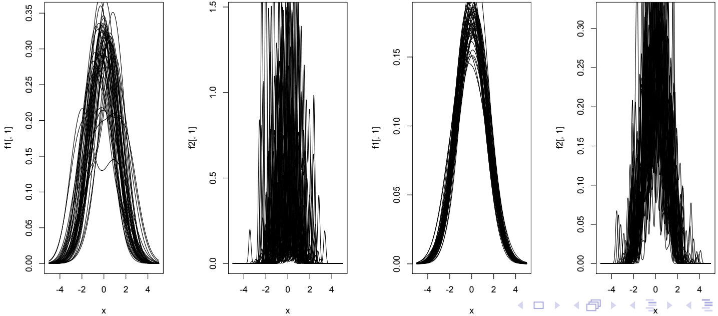
# Why mixtures ?

▶ **Mixtures of Gaussians**

$$K_\sigma P(x) = \int_{\mathbb{R}^d} \phi_\sigma(x-\mu) dP(\mu), \quad P = \text{proba}$$

- Analytic
- If $f_0$ *ordinary smooth* left : M=5 & sigma=1, sigma=0.1– Right : M=100 & sigma=1, sigma=0.1

$$K_\sigma f_0 \to f_0, \quad \sigma \to 0, \quad \& \exists P \quad K_\sigma P \approx K_\sigma f_0$$



# Remarks – towards asymptotic properties

- Can we assess the impact of hyperparameters ?

# Remarks – towards asymptotic properties

- Can we assess the impact of hyperparameters ?
- Are some hyperparameters more influencial than others ?

# Remarks – towards asymptotic properties

- Can we assess the impact of hyperparameters ?
- Are some hyperparameters more influencial than others ?
- Understand how the prior model acts as an approximation tool for the curve of interest ?

# A short history : from consistency to Bernstein von Mises and frequentist coverage

- First results : consistency until late 90s
  *L. Schwartz, A. Barron, Wasserman*
  Generic method to prove consistency

# A short history : from consistency to Bernstein von Mises and frequentist coverage

- First results : consistency until late 90s
  *L. Schwartz, A. Barron, Wasserman*
  Generic method to prove consistency
- Posterior concentration rates : 2000-
  *Ghosal, Ghosh & van der Vaart + children*
  • Generic method to obtain rates

minimax (adaptive ) Bayesian nonparametric estimators

  • Extension to empirical Bayes *Donnet et al., R. & Szabo*

# A short history : from consistency to Bernstein von Mises and frequentist coverage

- First results : consistency until late 90s
  *L. Schwartz, A. Barron, Wasserman*
  Generic method to prove consistency
- Posterior concentration rates : 2000-
  *Ghosal, Ghosh & van der Vaart + children*
  - Generic method to obtain rates

  minimax (adaptive ) Bayesian nonparametric estimators
  - Extension to empirical Bayes *Donnet et al., R. & Szabo*
- Semi & non - parametrics : More precise : BVM 2010 -
  *Castillo, Rivoirard & R., Bontemps, Kruijer, Kleijn, Castillo & Nickl, Spokoiny*
  some positive and negative (biais)

# A short history : from consistency to Bernstein von Mises and frequentist coverage

- First results : consistency until late 90s
  *L. Schwartz, A. Barron, Wasserman*
  Generic method to prove consistency
- Posterior concentration rates : 2000-
  *Ghosal, Ghosh & van der Vaart + children*
  • Generic method to obtain rates
  minimax (adaptive ) Bayesian nonparametric estimators
  • Extension to empirical Bayes *Donnet et al., R. & Szabo*
- Semi & non - parametrics : More precise : BVM 2010 -
  *Castillo, Rivoirard & R., Bontemps, Kruijer, Kleijn, Castillo & Nickl, Spokoiny*
  some positive and negative (biais)
- in between : (freq.) coverage & understanding posterior concentration 2013 -
  *VdV et al. ; Hoffman, R. , Schmidt-Hieger*

# Outline

# Posterior consistency and concentration rates

$$X^n = (X_1, ..., X_n) \sim P_\theta, \theta \in \Theta, \quad \theta \sim \Pi$$

- **Consistency** $d(\theta_1, \theta_2)$ = distance (or loss), $\theta_0 \in \Theta$
*the posterior is consistent* at $\theta_0$ iff $\forall \epsilon > 0$ $P_{\theta_0}$ a.s. or in proba.

$$\Pi[A_\epsilon | X^n] = 1 + o(1), \quad A_\epsilon = \{\theta \in \Theta; d(\theta_0, \theta) < \epsilon\}$$

- **Concentration rates** *the posterior concentrates* at the rate at least $\epsilon_n$ at $\theta_0$ iff

$$E^n_{\theta_0}[\Pi[A_{\epsilon_n} | X^n]] = 1 + o(1), \quad \epsilon_n \downarrow 0$$

• It depends on $d(.,.)$ and on $\Pi$ and $\theta_0$

- **Minimax concentration rates**

$$\sup_{\theta_0 \in \Theta_0} E^n_{\theta_0}[\Pi[A_{\epsilon_n} | X^n]] = 1 + o(1), \quad \epsilon_n \downarrow 0$$

with $\epsilon_n =$ minimax rate associated to $d(.,.)$ over $\Theta_0$ .

# Concentration rates : Ghosal &Van der Vaart- $A_{\epsilon_n} = \{\theta; d(\theta_0, \theta) \leq \epsilon_n\}$

$$P^{\pi}\left[A^c_{\epsilon_n} | X^n\right] = \frac{\int_{A^c_{\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)} d\pi(\theta)}{\int_{\Theta} e^{\ell_n(\theta)-\ell_n(\theta_0)} d\pi(\theta)} = o_{p_{\theta_0}}(1)??$$

- **Kullback-Leibler support condition :**

$$\pi(S_n) \geq e^{-c_1 n {\epsilon_n}^2}, \quad S_n = \{\theta; KL(f^n_{\theta_0}, f^n_\theta) \leq n{\epsilon_n}^2; V(f^n_{\theta_0}, f^n_\theta) \leq n{\epsilon_n}^2\}$$

# Concentration rates : Ghosal &Van der Vaart- $A_{\epsilon_n} = \{\theta; d(\theta_0, \theta) \le \epsilon_n\}$

$$P^\pi \left[A^c_{\epsilon_n} | X^n\right] = \frac{\int_{A^c_{\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)} d\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)-\ell_n(\theta_0)} d\pi(\theta)} = o_{p_{\theta_0}}(1)??$$

- **Kullback-Leibler support condition :**

$$\pi(S_n) \ge e^{-c_1 n {\epsilon_n}^2}, \quad S_n = \{\theta; KL(f^n_{\theta_0}, f^n_\theta) \le n{\epsilon_n}^2; V(f^n_{\theta_0}, f^n_\theta) \le n{\epsilon_n}^2\}$$

- **Sieves :** $\exists \Theta_n \subset \Theta$,

$$\pi(\Theta^c_n) = o(e^{-(c_1+1)n\epsilon_n^2})$$

# Concentration rates : Ghosal &Van der Vaart- $A_{\epsilon_n} = \{\theta; d(\theta_0, \theta) \leq \epsilon_n\}$

$$P^\pi \left[A_{\epsilon_n}^c | X^n\right] = \frac{\int_{A_{\epsilon_n}^c} e^{\ell_n(\theta) - \ell_n(\theta_0)} d\pi(\theta)}{\int_\Theta e^{\ell_n(\theta) - \ell_n(\theta_0)} d\pi(\theta)} = o_{p_{\theta_0}}(1)??$$

- **Kullback-Leibler support condition :**

$$\pi(S_n) \geq e^{-c_1 n {\epsilon_n}^2}, \quad S_n = \{\theta; KL(f_{\theta_0}^n, f_\theta^n) \leq n{\epsilon_n}^2; V(f_{\theta_0}^n, f_\theta^n) \leq n{\epsilon_n}^2\}$$

- **Sieves :** $\exists \Theta_n \subset \Theta$,

$$\pi(\Theta_n^c) = o(e^{-(c_1+1)n\epsilon_n^2})$$

- **Tests on** $\Theta_n$, $\exists \phi_n(\mathbf{x}^n) \in [0,1]$

$$E_{\theta_0}[\phi_n] = o(1), \qquad \sup_{\theta \in \Theta_n, d(\theta,\theta_0) > \epsilon_n} E_\theta[1 - \phi_n] = o(e^{-(c_1+3)n\epsilon_n^2})$$

# Proof of Ghosal & VdV.

$$A^c_{\epsilon_n} = \{d(\theta,\theta_0) > M\epsilon_n\},\ S_n = \{K_n(\theta_0,\theta) \le n\epsilon_n^2;\ V(\theta_0,\theta) \le n\epsilon_n^2\}$$

$$E_{\theta_0}\left[\Pi\left(A^c_{\epsilon_n}|X^n\right)\right] = E_{\theta_0}\left[\frac{\int_{A^c_{\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)}d\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)-\ell_n(\theta_0)}d\pi(\theta)}\right] := E_{\theta_0}\left[\frac{N_n}{D_n}\right]$$

$$\le E_{\theta_0}[\phi_n] + P^n_{\theta_0}\left[D_n < e^{-2n\epsilon_n^2}\pi(S_n)\right]^{(**)}$$

$$+ \frac{e^{2n\epsilon_n^2}}{\pi(S_n)} E^n_{\theta_0}[N_n(1-\phi_n)]$$

$$\le E_{\theta_0}[\phi_n] + \frac{\int_{S_n} P_{\theta_0}\left[\ell_n(\theta)-\ell_n(\theta_0) < -2n\epsilon_n^2\right]d\pi(\theta)}{\pi(S_n)}$$

$$+ \frac{e^{2n\epsilon_n^2}}{\pi(S_n)}\int_{A^c_{\epsilon_n}\cap\Theta_n} E_\theta[1-\phi_n]d\pi(\theta) + \frac{e^{2n\epsilon_n^2}}{\pi(S_n)}\Pi(\Theta_n^c)$$

# Lower bound on $D_n$

$$D_n \geq \int_{S_n} e^{\ell_n(\theta)-\ell_n(\theta_0)} d\pi(\theta)$$
$$\geq e^{-2n\epsilon^2} \int_{S_n} \mathbb{I}_{\ell_n(\theta)-\ell_n(\theta_0)\geq -2n\epsilon_n^2} d\pi(\theta)$$

So that

$$P_{\theta_0}^n \left[ D_n < e^{-2n\epsilon_n^2}\pi(S_n)/2 \right] \leq P_{\theta_0}^n \left[ \int_{S_n} \mathbb{I}_{\ell_n(\theta)-\ell_n(\theta_0)\leq -2n\epsilon_n^2} d\pi(\theta) \leq \pi(S_n)/ \right.$$
$$\leq \frac{2 \int_{S_n} P_{\theta_0} \left( \ell_n(\theta) - \ell_n(\theta_0) \leq -2n\epsilon_n^2 \right) d\pi(\theta)}{\pi(S_n)}$$

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.

- **Various families of priors**

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large $p$ small $n$ linear regression, covariance estimation, completion matrix, . . .

- **Various families of priors**

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large $p$ small $n$ linear regression, covariance estimation, completion matrix, . . .
  - Other complex models : Non parametric mixture models and NP HMMs : unknown emission distribution
- **Various families of priors**

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large *p* small *n* linear regression, covariance estimation, completion matrix, . . .
  - Other complex models : Non parametric mixture models and NP HMMs : unknown emission distribution
- **Various families of priors**
  - Gaussian processes, bases expansions

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large $p$ small $n$ linear regression, covariance estimation, completion matrix, . . .
  - Other complex models : Non parametric mixture models and NP HMMs : unknown emission distribution
- **Various families of priors**
  - Gaussian processes, bases expansions
  - Mixture models

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large $p$ small $n$ linear regression, covariance estimation, completion matrix, . . .
  - Other complex models : Non parametric mixture models and NP HMMs : unknown emission distribution
- **Various families of priors**
  - Gaussian processes, bases expansions
  - Mixture models
  - Polya trees

Hierarchical modelling for adaptation

# Applications to a wide class of problems

- **Various models**
  - Standard nonparametrics : nonlinear regression, density, classification, inverse problems etc.
  - High dimensional models : large $p$ small $n$ linear regression, covariance estimation, completion matrix, . . .
  - Other complex models : Non parametric mixture models and NP HMMs : unknown emission distribution
- **Various families of priors**
  - Gaussian processes, bases expansions
  - Mixture models
  - Polya trees
  - sparse models : spike and slab priors etc.

Hierarchical modelling for adaptation

# An example : NP Mixtures

- **Non parametric model for densities**

$$f_{P,\sigma}(x) = \int_{\mathbb{R}} \varphi_\sigma(x-\mu)dP(\mu)$$

• Discrete mixing distributions = $DP(M,G)$ or MFM

$$P = \sum_{j=1}^{K} p_j \delta_{(\mu_j)}, \quad K \sim \pi_K, \ (p_1,\cdots,p_K)|K \sim \pi_p \ \mu_j \overset{iid}{\sim} G$$

• prior on $\sigma$ : $\sigma \sim IG(a,b)$.

- **Result** If $\log f_0$ locally Hölder $\mathcal{H}_0(\alpha, L)$ $\alpha > 0$,

$(\mathbf{x}_1,\cdots,\mathbf{x}_n) \overset{iid}{\sim} f_0$

$$E^n_{f_0}\left[\Pi\left(\|f_0 - f_p\|_1 \lesssim n^{-\alpha/(2\alpha+1)}(\log n)^t|\mathbf{x}^n\right)\right] \to 1$$

Adaptive (over $\alpha$) minimax rate

# Empirical Bayes : data dependent prior

- **Setup** prior model $\Pi(d\theta|\lambda)$ , $\lambda \in \Lambda$
e.g. $\theta \in \mathbb{R}$, $\Pi(d\theta|\lambda) \equiv \mathcal{N}(\mu_0, \tau_0^2)$ & $\lambda = (\mu_0, \tau_0^2)$.
- **How to select $\lambda$ ?**
  - Prior information : informative prior

# Empirical Bayes : data dependent prior

- **Setup** prior model $\Pi(d\theta|\lambda)$ , $\lambda \in \Lambda$

e.g. $\theta \in \mathbb{R}$, $\Pi(d\theta|\lambda) \equiv \mathcal{N}(\mu_0, \tau_0^2)$ & $\lambda = (\mu_0, \tau_0^2)$.

- **How to select $\lambda$ ?**
  - Prior information : informative prior
  - Hierarchical $\lambda \sim Q$ : Hierarchical Bayes. But $Q$ ?

# Empirical Bayes : data dependent prior

- **Setup** prior model $\Pi(d\theta|\lambda)$ , $\lambda \in \Lambda$
e.g. $\theta \in \mathbb{R}$, $\Pi(d\theta|\lambda) \equiv \mathcal{N}(\mu_0, \tau_0^2)$ & $\lambda = (\mu_0, \tau_0^2)$.
- **How to select $\lambda$ ?**
  - Prior information : informative prior
  - Hierarchical $\lambda \sim Q$ : Hierarchical Bayes. But $Q$ ?
  - use data : $\hat{\lambda}(X^n)$ : empirical Bayes : double use of the data

# Examples of ways of choosing $\hat{\lambda}$ and examples

- **Maximum marginal likelihood estimate**

$$\hat{\lambda}_n = \text{argmax}_\lambda m(X^n|\lambda), \quad m(X^n|\lambda) = \int_\Theta f_\theta^n(X^n) d\Pi(\theta|\lambda)$$

- **Others** Moment - types estimate

$$X_1, \ldots, X_n|(F, \sigma) \overset{\text{i.i.d.}}{\sim} p_{F,\sigma}(\cdot) := \int \phi(\cdot|\mu, \sigma^2)\, \mathrm{d}F(\mu).$$

$$\theta = (F, \sigma), \quad \text{Prior} : F \sim DP(\alpha\mathcal{N}(\lambda, \tau^2)), \quad \sigma \sim \pi_\sigma$$

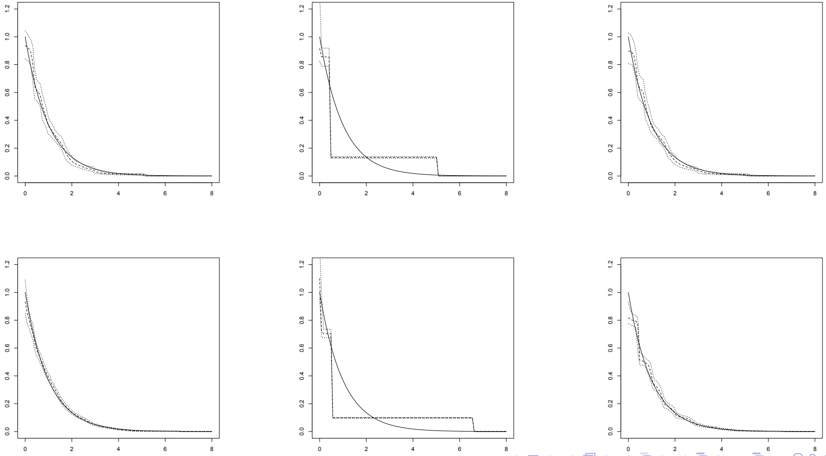
$$\hat{\lambda}_n = \bar{X}_n, \quad \hat{\tau}_n^2 = S_n^2, \max X_i - \min X_i$$

see e.g. Green & Richardson

# Outline

# Driving example : Poisson inhomogeneous monotone intensity estimation

Strategy 1 (Empirical) Strategy 2 ($\gamma$ fixed) Strategy 3 (hierarchica



# Dealing with data dependent priors

- Theory : so far fully Bayes
- How to adapt to data dependent priors ? ▶ **Ghosal and Van der Vaart 's proof** : Fubini

$$E_{\theta_0}\left[\Pi\left(U_n^c|X^n\right)\right] = E_{\theta_0}\left[\frac{\int_{A^c_{\epsilon_n}} e^{\ell_n(\theta)-\ell_n(\theta_0)}d\pi(\theta)}{\int_\Theta e^{\ell_n(\theta)-\ell_n(\theta_0)}d\pi(\theta)}\right] := E_{\theta_0}\left[\frac{N_n}{D_n}\right]$$

$$\leq E_{\theta_0}[\phi_n] + P^n_{\theta_0}\left[D_n < e^{-2n\epsilon_n^2}\pi(S_n)\right]$$

$$+ \frac{e^{2n\epsilon_n^2}}{\pi(S_n)} E^n_{\theta_0}[N_n(1-\phi_n)]$$

$$\leq E_{\theta_0}\left[\phi_n\right] + \frac{\int_{S_n} P_{\theta_0}\left[\ell_n(\theta)-\ell_n(\theta_0) < -2n\epsilon_n^2\right] d\pi(\theta)}{\pi(S_n)}$$

$$+ \frac{e^{2n\epsilon_n^2}}{\pi(S_n)}\int_{A^c_{\epsilon_n}\cap\Theta_n} E_\theta\left[1-\phi_n\right]d\pi(\theta) + \frac{e^{2n\epsilon_n^2}}{\pi(S_n)}\Pi(\Theta_n^c)$$

# Difficulty for $\pi\left(A^c_{\epsilon_n}|X^n;\hat\lambda\right) = o_p(1)$

- **If** $P_{\theta_0}\left[\hat\lambda_n \in \mathcal{K}_n\right] = 1 + o(1)$

$$\pi\left(A^c_{\epsilon_n}|X^n;\hat\lambda\right) \le \sup_{\lambda\in\mathcal{K}_n} \pi\left(A^c_{\epsilon_n}|X^n;\lambda\right) = o_p(1)?, \quad A_{\epsilon_n} = \{\theta, d(\theta_0,\theta) \le \epsilon_n\}$$

- **Non dominated models** $\lambda \to \Pi(d\theta|\lambda)$ : not dominated $\Rightarrow$ cannot study

$$\frac{\pi(\theta|\lambda)}{\pi(\theta|\lambda')}$$

# Outline

# Change of measure - $\sup_{\lambda \in \mathcal{K}_n} \pi(B_n | X^n; \lambda) = o_p(1)$?

- **A key tool** For all $\lambda, \lambda'$

$$\theta \sim \pi(\cdot|\lambda) \Rightarrow \psi_{\lambda,\lambda'}(\theta) \sim \pi(\cdot|\lambda')$$

- **Important class of examples** Mixtures (parametric or NP)

$\theta = (P, \phi)$

$$f_{P,\phi}(x) = \int K_\phi(x|z) dP(z) = \sum_j p_j K_\phi(x|z_j), \; P \sim DP(MG(\cdot|\lambda)), \; \phi \sim \pi_\phi$$

$$\psi_{\lambda,\lambda'}(f_{P,\phi})(x) = \sum_{j=1}^{\infty} p_j K_\phi(x|G^{-1}(G(z_j|\lambda)|\lambda'))$$
$$= f_{P',\phi}, \quad P' \sim DP(M, G(\cdot|\lambda'))$$

# A general Theorem : Same types of conditions as G& VdV

$$\sup_{\lambda' \in \mathcal{K}_n} \pi(A_{\epsilon_n}^c | X^n \lambda') = \sup_{\lambda' \in \mathcal{K}_n} \frac{\int_{A_{\epsilon_n}^c} p_{\psi_{\lambda,\lambda'}(\theta)}^{(n)}(x^n) d\pi(\theta|\lambda)}{\int_{\Theta} p_{\psi_{\lambda,\lambda'}(\theta)}^{(n)}(x^n) d\pi(\theta|\lambda)} := \frac{N_n}{D_n} = o(1)$$

$$\mathcal{K}_n = \cup_{i=1}^{N_n(u_n)} B(\lambda_i, u_n) \quad \Rightarrow \sup_{\lambda \in \mathcal{K}_n} = \max_i \sup_{\lambda \in B(\lambda_i, u_n)}$$

- **KL support condition** :
  - Non data dependent priors :
    $\Pi(\{KL(f_{\theta_0}^n, f_\theta^n) \leq n\epsilon_n^2; V(f_{\theta_0}^n, f_\theta^n) \leq n\epsilon_n^2\}) \gtrsim e^{-cn\epsilon_n^2}$

# A general Theorem : Same types of conditions as G& VdV

$$\sup_{\lambda' \in \mathcal{K}_n} \pi(A^c_{\epsilon_n}|X^n \lambda') = \sup_{\lambda' \in \mathcal{K}_n} \frac{\int_{A^c_{\epsilon_n}} p^{(n)}_{\psi_{\lambda,\lambda'}(\theta)}(x^n) d\pi(\theta|\lambda)}{\int_\Theta p^{(n)}_{\psi_{\lambda,\lambda'}(\theta)}(x^n) d\pi(\theta|\lambda)} := \frac{N_n}{D_n} = o(1)$$

$$\mathcal{K}_n = \cup_{i=1}^{N_n(u_n)} B(\lambda_i, u_n) \quad \Rightarrow \sup_{\lambda \in \mathcal{K}_n} = \max_i \sup_{\lambda \in B(\lambda_i, u_n)}$$

- **KL support condition** :
  - Non data dependent priors :
    $\Pi(\{KL(f^n_{\theta_0}, f^n_\theta) \leq n\epsilon_n^2; V(f^n_{\theta_0}, f^n_\theta) \leq n\epsilon_n^2\}) \gtrsim e^{-cn\epsilon_n^2}$
  - Data dependent priors :

$$\sup_{\lambda \in \mathcal{K}_n} \sup_{\theta \in \tilde{B}_n} P^{(n)}_{\theta_0} \left\{ \inf_{\|\lambda' - \lambda\| \leq u_n} \ell_n(\psi_{\lambda,\lambda'}(\theta)) - \ell_n(\theta_0) < -n\epsilon_n^2 \right\} = o(N_n(u_n$$

$$\pi(\tilde{B}_n) \gtrsim e^{-cn\epsilon_n^2}$$

# A general Theorem : Same types of conditions as G& VdV - II

- **tests** : Let $dQ^{\theta}_{\lambda,n}(x) = \sup_{\|\lambda' - \lambda\| \leq u_n} p^{(n)}_{\psi_{\lambda,\lambda'}(\theta)}(x) d\mu(x)$,

$$E^{(n)}_{\theta_0}(\phi_n) = o(1), \quad \sup_{\lambda \in \mathcal{K}_n} \sup_{d(\theta,\theta_0) > \epsilon_n, \Theta_n} \int_{\mathcal{X}^n} (1 - \phi_n) dQ^{\theta}_{\lambda,n}(x^n) \leq e^{-Kn\epsilon_n^2}$$

$$\log N_n(u_n) = o(n\epsilon_n^2)$$

- $\Theta_n^c$
  - Non data dependent priors : $\pi(\Theta_n^c) \leq e^{-cn\epsilon_n^2}$

# A general Theorem : Same types of conditions as G& VdV - II

- **tests** : Let $dQ^{\theta}_{\lambda,n}(x) = \sup_{\|\lambda'-\lambda\|\leq u_n} p^{(n)}_{\psi_{\lambda,\lambda'}(\theta)}(x)d\mu(x)$,

$$E^{(n)}_{\theta_0}(\phi_n) = o(1), \quad \sup_{\lambda\in\mathcal{K}_n} \sup_{d(\theta,\theta_0)>\epsilon_n,\Theta_n} \int_{\mathcal{X}^n}(1-\phi_n)dQ^{\theta}_{\lambda,n}(x^n) \leq e^{-Kn\epsilon_n^2}$$

$$\log N_n(u_n) = o(n\epsilon_n^2)$$

- $\Theta_n^c$
  - Non data dependent priors : $\pi(\Theta_n^c) \leq e^{-cn\epsilon_n^2}$
  - Data dependent priors

$$\int_{\Theta_n^c} Q^{\theta}_{\lambda,n}(\mathcal{X}^n)\pi(d\theta|\lambda) \leq e^{-cn\epsilon_n^2}$$

# A general Theorem : comments

$$\pi\left(d(\theta,\theta_0) \leq \epsilon_n | x^n, \hat\lambda_n\right) = o_{p_0}(1)$$

- If $\mathcal{K}_n = \{\lambda; \epsilon_n(\lambda) \leq M_n\epsilon_n^*\}$, then

$$\epsilon_n \leq M_n\epsilon_n^*, \quad \epsilon_n^* = \inf\{\epsilon_n(\lambda); \lambda \in \Lambda\}$$

$$\Downarrow$$

Oracle posterior concentration rates

- BUT : need to know $\mathcal{K}_n$ e.g. MMLE R& Szabo (2015)

# Application to DP mixtures of Gaussians

- **Model** $x^n = (x_1, \cdots, x_n)$ iid $f$
- **prior on** $f$ **: DPM Gaussian**

$$f_{P,\sigma}(x) = \int_{\mathbb{R}} \phi_\sigma(x-\mu)dP(\mu), \quad P \sim DP(A\mathcal{N}(\mu_0, \tau^2)), \quad \sigma \sim \pi_\sigma$$

- **Choice for** $\mu_0, \tau^2$ **?** $\lambda = (\mu_0, \tau^2)$ Two cases :

$$\hat{\mu}_0 = \bar{x}_n, \quad \hat{\tau} = s_n, \quad \text{or } \hat{\mu}_0 = \bar{x}_n, \quad \hat{\tau} = \max_i x_i - \min_i x_i$$

- **Change of measure**

$$\psi_{\lambda,\lambda'}(f_P)(x) = \sum_{j=1}^{\infty} p_j \phi_\sigma(x - \mu_j + \Delta_j), \quad \Delta = \mu_j\left(\frac{\tau'}{\tau} - 1\right) - \mu_0\tau' + \mu_0'$$

Then

$$\psi_{\lambda,\lambda'}(f_P) \sim DPM(A\mathcal{N}(\mu_0', \tau')), \quad \text{when} \quad P \sim DP(A\mathcal{N}(\mu_0, \tau))$$

# Results for DP mixtures of Gaussians

$$f_{P,\sigma}(x) = \int_{\mathbb{R}} \phi_\sigma(x-\mu)dP(\mu), \quad P \sim DP(A\mathcal{N}(\mu_0,\tau^2)), \quad \sigma \sim \pi_\sigma$$

## Theorem

*Under same conditions as in fully Bayes* $\exists a > 0$ *such that if* $\mathcal{K}_n \subset [a_1, a_2] \times [\tau_1, (\log n)^q]$, *if* $f_0 \in \mathcal{H}_{\text{loc}}(\alpha)$

$$\pi\left(\|f_{P,\sigma} - f_0\|_1 > (\log n)^a n^{-\alpha/(2\alpha+1)} | \mathbf{x}^n\right) = o_{p_0}(1)$$

• Applies to $\hat{\lambda}_n = (\bar{x}_n, s_n)$ and $(\bar{x}_n, \max_i x_i - \min_i x_i)$ : in the latter loss in $\log n$

• $(\bar{x}_n, \max_i x_i - \min_i x_i)$ : acts like a non informative prior

# Some examples of transformations

- Gaussian processes

$$\sum_j \theta_j \phi_j, \quad \theta_j \overset{ind}{\sim} \mathcal{N}(0, \tau_j^2), \tau_j = \tau j^{-\alpha-1/2}$$

  - $\lambda = \tau$

$$\psi(\theta_j) = \frac{\tau'}{\tau}\theta_j$$

- Splines : $\sum_{j=1}^K \theta_j B_j, \quad \theta_j \overset{iid}{\sim} \tau g$

# Some examples of transformations

- Gaussian processes

$$\sum_j \theta_j \phi_j, \quad \theta_j \overset{ind}{\sim} \mathcal{N}(0, \tau_j^2), \tau_j = \tau j^{-\alpha-1/2}$$

  - $\lambda = \tau$

$$\psi(\theta_j) = \frac{\tau'}{\tau}\theta_j$$

  - $\lambda = \alpha$

$$\psi(\theta_j) = j^{\alpha-\alpha'}\theta_j$$

- Splines : $\sum_{j=1}^{K} \theta_j B_j, \quad \theta_j \overset{iid}{\sim} \tau g$

# Some examples of transformations

- Gaussian processes

$$\sum_j \theta_j \phi_j, \quad \theta_j \overset{ind}{\sim} \mathcal{N}(0, \tau_j^2), \tau_j = \tau j^{-\alpha-1/2}$$

  - $\lambda = \tau$

$$\psi(\theta_j) = \frac{\tau'}{\tau}\theta_j$$

  - $\lambda = \alpha$

$$\psi(\theta_j) = j^{\alpha-\alpha'}\theta_j$$

- Splines : $\sum_{j=1}^K \theta_j B_j, \quad \theta_j \overset{iid}{\sim} \tau g$
  - $K$ : no need for transformation

# Some examples of transformations

- Gaussian processes

$$\sum_j \theta_j \phi_j, \quad \theta_j \overset{ind}{\sim} \mathcal{N}(0, \tau_j^2), \tau_j = \tau j^{-\alpha-1/2}$$

  - $\lambda = \tau$

$$\psi(\theta_j) = \frac{\tau'}{\tau}\theta_j$$

  - $\lambda = \alpha$

$$\psi(\theta_j) = j^{\alpha-\alpha'}\theta_j$$

- Splines : $\sum_{j=1}^K \theta_j B_j, \quad \theta_j \overset{iid}{\sim} \tau g$
  - $K$ : no need for transformation
  - $\tau \rightarrow$

$$\psi_{\tau,\tau'}(\theta_j) = \frac{\tau'}{\tau}\theta_j$$

# Some important issues ?

- **Importance of the loss function**

General theory of G& VdV holds for *testable* losses :

$$E^n_{\theta_0}[\phi_n] = o(1), \qquad \sup_{\{d(\theta_0,\theta)>M\epsilon_n\}\cap\Theta_n} E^n_f(1-\phi_n) \leq e^{-(c+2)n\epsilon_n^2}$$

e.g. If $\theta_0 \in \Theta_0 = \mathcal{H}(\alpha, L)$ and $d(\theta_0,\theta) = \|\theta_0 - \theta\|_\infty \Rightarrow$ need for more involved approach

Important to explore various features of the posterior

- **Credible / Confidence statements**

If $C_\alpha$ is $\alpha$- credible : $\Pi(C_\alpha|X^n) = 1-\alpha$ We have

$$\int_\Theta P_\theta(\theta \in C_\alpha) d\Pi(\theta) = 1-\alpha$$

Posterior concentration $\epsilon_n \Rightarrow$ if $C_\alpha = \{d(\theta,\hat{\theta}) \leq q_\alpha(\mathbf{x})\}$ Then

$$E_\theta[|C_\alpha|] = O(\epsilon_n)$$

$$\text{But} \quad \inf_{\theta\in\Theta} P_\theta(\theta \in C_\alpha) \geq ??$$

- **Semi - parametric**

# On adaptive honest confidence regions (Low, Cai and Low)

- **Honest confidence regions**

$$\inf_{\theta\in\Theta} P_\theta\left(\theta \in C_n\right) \geq 1-\alpha$$

- **Adaptive confidence regions** : size

$$\sup_{\beta\in(\beta_1,\beta_2)} \sup_{f\in\mathcal{S}_\beta(L)} \epsilon_n(\beta)^{-1} E_\theta\left(|C_n|\right) = O(1)$$

- **Problem :** It essentially is not possible

# Approach of Nickl and al.

- Construct $C_n$ honest over $\Theta$

# Approach of Nickl and al.

- Construct $C_n$ honest over $\Theta$
- Consider $\tilde{\Theta} \subset \Theta$ of *good* points and define

$$\tilde{C}_n = C_n \cap \tilde{\Theta}$$

# Approach of Nickl and al.

- Construct $C_n$ honest over $\Theta$
- Consider $\tilde{\Theta} \subset \Theta$ of *good* points and define

$$\tilde{C}_n = C_n \cap \tilde{\Theta}$$

- Show honesty and size over $\tilde{\Theta}$

# Impact of posterior concentration rates on coverage : the nonparametric case . Castillo & Nickl, K. Ray

$$\text{if } E_{\theta_0}\left[\Pi\left(d(\theta_0,\theta) \le \epsilon_n | \mathbf{x}^n\right)\right] = 1 + o(1), \quad E_{\theta_0}\left[d(\hat\theta_n, \theta_0)\right] \lesssim \epsilon_n$$

then with $C_\alpha = \{\theta; d(\hat\theta_n, \theta) \le q_n(\alpha)\}$, and $\Pi(C_\alpha | \mathbf{x}^n) = 1 - \alpha$,

$$E_{\theta_0}\left[|C_\alpha|\right] \lesssim \epsilon_n, \quad \int_\Theta P_\theta(\theta \in C_\alpha)\Pi(d\theta) = 1 - \alpha$$

But typically (Cox)

$$\liminf_n P_\theta(\theta \in C_\alpha) = 0$$

# Use of weak BvM : Castillo & Nickl

- **Gaussian white noise model** On wavelet basis

$$X_{j,k} = \theta_{j,k} + n^{-1/2}\epsilon_{j,k}, \quad \epsilon_{j,k} \overset{iid}{\sim} \mathcal{N}(0,1)$$

- **Prior $\Pi$ on $\theta$** $w = (w_j)_{j\in\mathbb{N}}$ $w_j >> \sqrt{j}$

$$\beta_{\mathcal{M}_0(w)}(\Pi(\sqrt{n}(.-X)|\mathbf{x}^n), \mathcal{N}) = o_p(1)$$

$$\Downarrow$$

Credible ball

$$C_n = \{\theta; \sup_{j,k} |\frac{|\theta_{j,k} - X_{j,k}|}{w_j} \leq R_\alpha\}, \quad \Pi(C_n|\mathbf{x}^n) = 1-\alpha$$

Then

$$\inf_{\theta_0\in\tilde{\Theta}} P_{\theta_0}(\theta_0 \in C_n) \to 1-\alpha$$

- **Size ?** $\bar{C}_n = C_n \cap \{\theta; \|\theta\|_{\mathcal{H}(\gamma)} \leq u_n\}$

$$P_{\theta_0}(\theta_0 \in \bar{C}_n) \to 1-\alpha, \quad \theta_0 \in \mathcal{H}(\gamma)$$

# Limits (so far)

• Interpretation ? non usual norm (geometry)
• Extension outside $L_2$ structure ?
• Implies Bernstein von Mises of smooth functionals of $\theta \Rightarrow$
Good confidence of cerdible regions for such functionals

# Other approach : Szabo et al., van der Vaart

- **Gaussian white noise model**

$$X_j = \theta_j + n^{-1/2}\epsilon_j, \quad \epsilon_j \overset{iid}{\sim} \mathcal{N}(0,1)$$

- **Gaussian prior**

$$\theta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2 j^{-2\lambda-1})$$

- **empirical Bayes**

$$\hat{\lambda} = \sup_{\lambda} m_n(\lambda), \quad m_n(\lambda) = \int_{\Theta} L_n(\theta) d\pi_\lambda(\theta)$$

$$\Downarrow$$

$$C_n = \{\|\theta - \theta_0\|_2 \leq L_n q_n(\alpha)\}, \quad L_n \uparrow +\infty$$

- $$\inf_{\tilde{\Theta}} P_\theta\left(\theta \in C_n\right) \to 1$$

- **Gaussian white noise model**

$$X_j = \theta_j + n^{-1/2}\epsilon_j, \quad \epsilon_j \overset{iid}{\sim} \mathcal{N}(0,1)$$

- **Gaussian prior**

$$\theta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2 j^{-2\lambda-1})$$

- **empirical Bayes**

$$\hat{\lambda} = \sup_{\lambda} m_n(\lambda), \quad m_n(\lambda) = \int_\Theta L_n(\theta) d\pi_\lambda(\theta)$$

$$\Downarrow$$

$$C_n = \{\|\theta - \theta_0\|_2 \leq L_n q_n(\alpha)\}, \quad L_n \uparrow +\infty$$

- $$\inf_{\tilde{\Theta}} P_\theta\left(\theta \in C_n\right) \to 1$$
- Minimax adaptive size : $E_{\theta_0}|C_n| \lesssim \epsilon_n(\beta), \theta_0 \in \mathcal{S}_\beta$

# What is $\tilde{\Theta}$ ? : Polished tail condition

There exists $N_0 \geq 0$ $L_0 > 0$ s.t. $\forall N \geq N_0$

$$\sum_{i=N}^{\infty} \theta_i^2 \leq L_0 \sum_{i=N}^{\rho N} \theta_i^2$$

Regularly decreasing coefficients

# Extensions to more general models and priors R. & Szabo

- **Model** $f_\theta(X^n)$

$$\theta \in \Theta = \cup_{k\in\mathbb{N}^*}\Theta(k), \quad \dim(\Theta(k)) = d_k \asymp k$$

- **Prior**

$$k \sim \pi_k, \quad [\theta|k] \sim \pi_{|k}(\cdot)$$

- **Metric on** $\Theta$ $d(\theta_1, \theta_2)$ (*natural*)
  - Density $d(f_{\theta_1}, f_{\theta_2}) =$ Hellinger

# Extensions to more general models and priors R. & Szabo

- **Model** $f_\theta(X^n)$

$$\theta \in \Theta = \cup_{k\in\mathbb{N}^*}\Theta(k), \quad \dim(\Theta(k)) = d_k \asymp k$$

- **Prior**

$$k \sim \pi_k, \quad [\theta|k] \sim \pi_{|k}(\cdot)$$

- **Metric on** $\Theta$ $d(\theta_1, \theta_2)$ (*natural*)
  - Density $d(f_{\theta_1}, f_{\theta_2}) =$ Hellinger
  - White noise $L_2$

# Extensions to more general models and priors R. & Szabo

- **Model** $f_\theta(X^n)$

$$\theta \in \Theta = \cup_{k\in\mathbb{N}^*}\Theta(k), \quad \dim(\Theta(k)) = d_k \asymp k$$

- **Prior**

$$k \sim \pi_k, \quad [\theta|k] \sim \pi_{|k}(\cdot)$$

- **Metric on $\Theta$** $d(\theta_1, \theta_2)$ (*natural*)
  - Density $d(f_{\theta_1}, f_{\theta_2}) =$ Hellinger
  - White noise $L_2$
  - Regression $x_i = f(z_i) + \epsilon_i$,

$$d_n^2(f_{\theta_1}, f_{\theta_2}) = n^{-1}(\sum_i (f_{\theta_1}(z_i) - f_{\theta_2}(z_i))^2)$$

## Extensions to more general models and priors R. & Szabo

- **Model** $f_\theta(X^n)$

$$\theta \in \Theta = \cup_{k\in\mathbb{N}^*}\Theta(k), \quad \dim(\Theta(k)) = d_k \asymp k$$

- **Prior**

$$k \sim \pi_k, \quad [\theta|k] \sim \pi_{|k}(\cdot)$$

- **Metric on** $\Theta$ $d(\theta_1,\theta_2)$ (*natural*)
  - Density $d(f_{\theta_1}, f_{\theta_2})$ = Hellinger
  - White noise $L_2$
  - Regression $x_i = f(z_i) + \epsilon_i$,

$$d_n^2(f_{\theta_1}, f_{\theta_2}) = n^{-1}(\sum_i (f_{\theta_1}(z_i) - f_{\theta_2}(z_i))^2)$$

  - Classification : $P[x_i = 1|z_i] = q(z_i)$. Hellinger

$$d_n^2(q_{\theta_1}, q_{\theta_1}) = n^{-1}\sum_i h_n^2(q_{\theta_1}(z_i), q_{\theta_2}(z_i))$$

# Polished tail condition for $\theta_0$ and result

- **Bias**

$$b^2(k) = \inf_{\theta \in \Theta(k)} d^2(\theta_0, \theta)$$

- **Polished tail** $\exists R > 1, 1 > \tau > 0$

$$b(Rk) \leq \tau b(k), \quad \forall k \geq k_0$$

- **Together with other regularity conditions** : same result as in White noise

# Outline

# Semi-parametric Bayesian methods : setup

- **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 ; partial linear regression, semi - parametric HMMs, mixtures

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

# Semi-parametric Bayesian methods : setup

- **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 ; partial linear regression, semi - parametric HMMs, mixtures

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

  - $\theta =$ curve $f$, (density, regression, spectral density)

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

ex : $\psi(f) = F(x) = \int \mathbb{1}_{u \le x} f(t) dt$, $\psi(f) = \int f^2(u) du$, $\psi(f) = f(x_0)$

# Marginal posterior

$$\Pi(\psi(\theta) \in A_n | X^n)??$$

- **Regular models**

$$\exists \hat{\psi}, \text{ s.t. } \sqrt{n}(\hat{\psi} - \psi(\theta_0)) \to \mathcal{N}(0, v_0)$$

What about Bayesian approaches ?

$$\Pi(d(\psi, \psi(\theta_0)) \leq M_n n^{-1/2} | X^n) \to 1, \quad \forall M_n \uparrow +\infty?$$

More ? : asymptotic normality : BvM

$$\Pi(\sqrt{n}(\psi - \hat{\psi}) \in A | X^n) \to \mathbb{P}(\mathcal{N}(0, v_0) \in A)?$$

# 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 $\theta$ close to a Normal with mean $\hat{\theta}$ and variance $V_{\theta_0}(\hat{\theta})$ under $P_{\theta_0}$.

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

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

# illustration :

$X_i \sim P(\lambda)$, and $\pi(\lambda) = \Gamma(a, b)$ then

$$\pi(\lambda|X^n) = \Gamma(a+n\bar{X}_n, b+n), \quad a = 10, b = 1, \quad \lambda_0 = 1, \quad n = 1, 10, 100$$

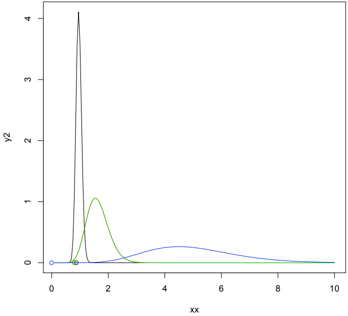



FIG.: posterior, n=1 = blue, n=10=green, n=100=black.

# Outline

# Applications of BVM

1. Construction of HPD regions

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

Then

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

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. $\alpha$ credible regions $C_\alpha^\pi$ for $\theta$ are asymptotically $\alpha$-confidence regions

$$P_\theta[\theta \in C_\alpha^\pi] = \alpha + o(1)$$

# 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. $\alpha$ credible regions $C_\alpha^\pi$ for $\theta$ are asymptotically $\alpha$-confidence regions

$$P_\theta[\theta \in C_\alpha^\pi] = \alpha + o(1)$$

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* $\forall \epsilon > 0$,

$$\lim_{M \to \infty} limsup_n P^\pi \left[ |\theta - \theta_0| > \epsilon | X^n \right] = 0,$$

*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* $\forall\epsilon > 0$,

$$\lim_{M\to\infty} limsup_n P^\pi\left[|\theta-\theta_0|>\epsilon|X^n\right]=0,$$

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

*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 ?

- **Localize** Since $P^\pi\left[|\theta - \theta_0| > \epsilon | X^n\right]$ : only need look at

$$|\theta - \theta_0| = o(1)$$

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

$$l_n(\theta) = \log f(X^n|\theta), \quad \hat\theta = \text{ post mean or normalized score}$$

$$\begin{aligned}\pi(\theta|X^n) &\propto e^{l_n(\theta) - l_n(\hat\theta) + \log(\pi(\theta)) - \log(\pi(\hat\theta)))} \\ &\propto e^{-\frac{(\theta-\hat\theta)J_n(\theta-\hat\theta)}{2}(1+o_P(1))} \quad \text{when } |\theta - \hat\theta| = o_P(1)\end{aligned}$$

$$J_n = D^2 l_n(\theta)|_{\theta=\hat\theta}$$

# Extension to nonparametric models

- **Control of the LAN rest** uniformly compared $n\|\theta - \theta_0\|_2^2$
- **Continuity of the prior density**

# Outline

# General BVM theorem : framework

- **Model :** $X^n|\theta \sim f_\theta^n$ where $\theta \in \Theta$ infinite dimensional
  $\pi$ : prior on $\theta$
- **Parameter of interest :** $\psi(\theta)$
- **Aim :** Asymptotic posterior distribution of $\psi(\theta)$ :
  - Normality ?

# General BVM theorem : framework

- **Model :** $X^n|\theta \sim f_\theta^n$ where $\theta \in \Theta$ infinite dimensional
  $\pi$ : prior on $\theta$
- **Parameter of interest :** $\psi(\theta)$
- **Aim :** Asymptotic posterior distribution of $\psi(\theta)$ :
  - Normality ?
  - Centering ? Variance ?

# Outline

- **LAN condition** $f_0^n = f_{\theta_0}^n$ (truth)

$$\log f_\theta^n(X^n) - \log f_0^n(X^n) = -\frac{n\|\theta - \theta_0\|_L^2}{2} + \sqrt{n} W_n(\theta - \theta_0) + R_n(\theta, \theta_0)$$

with $W_n(u) \sim \mathcal{N}(0, \|u\|_L^2)$ and $u \to W_n(u)$ linear.

- **Concentration** : $\exists A_n \subset \Theta \; P^\pi \left[A_n | X^n\right] = 1 + o_p(1)$ typically

$$A_n \subset \{d(\theta_0, \theta) \le \epsilon_n\}, \quad \epsilon_n \downarrow 0$$

- **Smoothness of** $\psi$

$$\psi(\theta) = \psi(\theta_0) + < \theta - \theta_0, \dot{\psi}_0 >_L + < \theta - \theta_0, \ddot{\psi}_0(\theta - \theta_0) >_L + r(\theta, \theta_0)$$

2 regimes

- Linear : $\ddot{\psi}_0 = 0$ — Here only this one

- **LAN condition** $f_0^n = f_{\theta_0}^n$ (truth)

$$\log f_\theta^n(X^n) - \log f_0^n(X^n) = -\frac{n\|\theta - \theta_0\|_L^2}{2} + \sqrt{n} W_n(\theta - \theta_0) + R_n(\theta, \theta_0)$$

with $W_n(u) \sim \mathcal{N}(0, \|u\|_L^2)$ and $u \to W_n(u)$ linear.

- **Concentration** : $\exists A_n \subset \Theta$ $P^\pi[A_n|X^n] = 1 + o_p(1)$ typically

$$A_n \subset \{d(\theta_0, \theta) \le \epsilon_n\}, \quad \epsilon_n \downarrow 0$$

- **Smoothness of** $\psi$

$$\psi(\theta) = \psi(\theta_0) + < \theta - \theta_0, \dot\psi_0 >_L + < \theta - \theta_0, \ddot\psi_0(\theta - \theta_0) >_L + r(\theta, \theta_0)$$

2 regimes

- Linear : $\ddot\psi_0 = 0$ — Here only this one
- quadratic $\ddot\psi_0 \neq 0$

# Examples of LAN

$$\log f_\theta^n(X^n) - \log f_0^n(X^n) = -\frac{n\|\theta - \theta_0\|_L^2}{2} + \sqrt{n} W_n(\theta - \theta_0) + R_n(\theta, \theta_0)$$

• White noise $dX(t) = f(t)dt + dW(t)/\sqrt{n}$ $(\Leftrightarrow X_i = \theta_i + n^{-1/2}\epsilon_i$, $i \in \mathbb{N})$

$$\ell_n(\theta) - \ell_n(\theta_0) = \frac{-n\|\theta - \theta_0\|_2^2}{2} + \sqrt{n}\sum_i (\theta_i - \theta_{0i})\epsilon_i$$

$$\|\theta - \theta_0\|_L^2 = \sum_{i=1}^{\infty} (\theta_i - \theta_{0i})^2$$

# LAN condition, Ex 2

- Density $X_i \sim f$ i.i.d $\theta = \log f$

$$\ell_n(\theta) - \ell_n(\theta_0) = \sum_i \theta(X_i) - \theta_0(X_i) = -\frac{n\|\theta - \theta_0\|_L^2}{2} + \sqrt{n}\mathbb{G}_n(\theta - \theta_0) + R_n(\theta)$$

$$\|\theta - \theta_0\|_L^2 = \int f_0(x)\left(\log f(x) - \log f_0(x)\right)^2 dx - \left(\int f_0(\log f - \log f_0)\right)^2$$

- auto- regression $Y_i = f(Y_{i-1}) + \epsilon_i$ , $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\|\theta - \theta_0\|_L^2 = \int_{\mathbb{R}} q_{f_0}(x)(f(x) - f_0(x))^2 dx$$

# Outline

# Theorem

Under LAN+Concentration+smooth if

$$\theta_t = \theta - t\frac{\dot{\psi}_0}{\sqrt{n}}, \quad t \neq 0$$

If on $A_n$, $R(\theta, \theta_0) - R(\theta_t, \theta_0) + t\sqrt{n} r(\theta, \theta_0) = o(1)$ and
- **The condition**

$$\frac{\int_{A_n} p_{\theta_t}(Y^n) d\pi(\theta)}{\int_{A_n} p_{\theta}(Y^n) d\pi(\theta)} = 1 + o_p(1)$$

Then a posteriori :

$$\sqrt{n}(\psi(\theta) - \hat{\psi}) \approx \mathcal{N}(0, V_{0,n}), \quad \hat{\psi} = \psi(\theta_0) + \frac{W_n(\dot{\psi}_0)}{\sqrt{n}}$$

$$V_{0,n} = \|\dot{\psi}_0\|_L^2$$

# Comments

- **LAN+ Concentration + smoothness** Usual type of condition. Posterior concentration rates (LAN norm)

# Comments

- **LAN+ Concentration + smoothness** Usual type of condition. Posterior concentration rates (LAN norm)
- **The condition** Means that we can consider a *change of parameters*

$$\theta_t = \theta - t\frac{\dot\psi_0}{\sqrt{n}}, \quad s.t.$$

$$d\pi(\theta_t) = d\pi(\theta)(1 + o(1))$$

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 $\pi$.

# linear regime : $\ddot{\psi}_0 = 0$

$$\theta_t = \theta_0 - t\frac{\dot{\psi}_0}{\sqrt{n}}$$

$$\& \quad \frac{\int_{A_n} p_{\theta_t}(Y^n) d\pi(\theta)}{\int_{A_n} p_{\theta}(Y^n) d\pi(\theta)} = 1 + o_p(1)$$

Then a posteriori :

$$\sqrt{n}(\psi(\theta) - \hat{\psi}) \approx \mathcal{N}(0, V_{0,n}), \quad \hat{\psi} = \psi(\theta_0) + \frac{W_n(\dot{\psi}_0)}{\sqrt{n}}$$

$$V_{0,n} = \|\dot{\psi}_0\|_L^2$$

BvM

# Example in linear regime

- **Model** $X_1, ..., X_n | f \sim f$ i.i.d $X_i \in [0, 1]$, $\theta = \log f$
- **functionals**
  - Entropy $\psi(f) = \int_0^1 f \log f(x) dx$ & $f$ smooth

$$\dot{\psi}_0 = \log f_0 - \psi(f_0), \quad \ddot{\psi}_0 = 0$$

- **Prior model** random histogram

$$f(x) = \sum_{j=1}^{k} \mathbb{I}_{I_j}(x) k w_j, \quad \sum w_j = 1, \quad I_j = ((j-1)/k, j/k]$$

$$(w_1, \cdots, w_k) \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$$

# Example in linear regime

- **Model** $X_1, ..., X_n | f \sim f$ i.i.d $X_i \in [0,1]$, $\theta = \log f$
- **functionals**
  - Entropy $\psi(f) = \int_0^1 f \log f(x) dx$ & $f$ smooth

$$\dot{\psi}_0 = \log f_0 - \psi(f_0), \quad \ddot{\psi}_0 = 0$$

  - Linear $\psi(f) = \int a(x) f(x) dx$.

$$\dot{\psi}_0 = a - \psi(f_0), \quad \ddot{\psi}_0 = 0$$

- **Prior model** random histogram

$$f(x) = \sum_{j=1}^{k} \mathbb{I}_{I_j}(x) k w_j, \quad \sum w_j = 1, \quad I_j = ((j-1)/k, j/k]$$

$$(w_1, \cdots, w_k) \sim \mathcal{D}(\alpha_1, \cdots, \alpha_k)$$

# Results

$$f_0 \in \mathcal{H}(\beta), \beta > 0, \quad \|\log f_0\|_\infty < +\infty$$

$$\begin{aligned}\theta_t &= \log f_{w,k} - \frac{t\dot{\psi}_0}{\sqrt{n}} = \log f_{w,k} - \frac{t\dot{\psi}_{[k]}}{\sqrt{n}} + \frac{t}{\sqrt{n}}[\dot{\psi}_{[k]} - \dot{\psi}_0] \\ &:= \theta_{t[k]} + \frac{t}{\sqrt{n}}[\dot{\psi}_{[k]} - \dot{\psi}_0], \quad w_j \to w_j - t\psi_j/\sqrt{n}, \, j \leq k\end{aligned}$$

and $A_{n,k} = \{f_{w,k}; h(f_{w,k}, f_{0[k]}) \lesssim \sqrt{k \log n / n}\}$

$$\ell_n(\theta_t) - \ell_n(\theta_{t[k]}) = \sqrt{n} \int (\dot{\psi}_{[k]} - \dot{\psi}_0)(f_{0[k]} - f_0) + \mathbb{G}_n(\dot{\psi}_{[k]} - \dot{\psi}_0) + o_p(1)$$

True for any $k \lesssim n/(\log n)^2$.

# Examples of functionals

$$\sqrt{n}\int(\dot\psi_{[k]} - \dot\psi_0)(f_{0[k]} - f_0) + \mathbb{G}_n(\dot\psi_{[k]} - \dot\psi_0)$$

- **Deterministic *k* case** : $k = K_n = \lfloor \sqrt{n}(\log n)^{-2} \rfloor$
  - Entropy : $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM Model $k$
- **random *k* case** : $k \sim \mathcal{P}(\lambda)$

# Examples of functionals

$$\sqrt{n}\int(\dot\psi_{[k]}-\dot\psi_0)(f_{0[k]}-f_0)+\mathbb{G}_n(\dot\psi_{[k]}-\dot\psi_0)$$

- **Deterministic *k* case** : $k = K_n = \lfloor \sqrt{n}(\log n)^{-2} \rfloor$
  - Entropy : $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM Model $k$
  - Linear & $a \in \mathcal{H}(\gamma)$, $\beta + \gamma > 1$ : BVM linear CDF : BVM
- **random *k* case** : $k \sim \mathcal{P}(\lambda)$

# Examples of functionals

$$\sqrt{n}\int(\dot\psi_{[k]}-\dot\psi_0)(f_{0[k]}-f_0)+\mathbb{G}_n(\dot\psi_{[k]}-\dot\psi_0)$$

- **Deterministic $k$ case** : $k = K_n = \lfloor \sqrt{n}(\log n)^{-2} \rfloor$
  - Entropy : $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM Model $k$
  - Linear & $a \in \mathcal{H}(\gamma)$, $\beta + \gamma > 1$ : BVM linear CDF : BVM
- **random $k$ case** : $k \sim \mathcal{P}(\lambda)$
  - entropy $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM

# Examples of functionals

$$\sqrt{n}\int(\dot\psi_{[k]}-\dot\psi_0)(f_{0[k]}-f_0)+\mathbb{G}_n(\dot\psi_{[k]}-\dot\psi_0)$$

- **Deterministic $k$ case** : $k = K_n = \lfloor \sqrt{n}(\log n)^{-2} \rfloor$
  - Entropy : $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM Model $k$
  - Linear & $a \in \mathcal{H}(\gamma)$, $\beta + \gamma > 1$ : BVM linear CDF : BVM
- **random $k$ case** : $k \sim \mathcal{P}(\lambda)$
  - entropy $\dot\psi = \log f_0 - \psi(f_0)$, $\beta > 1/2$ : BVM
  - Linear : Risk of bias : There are counterexamples

# Outline

# Weak BvM for smooth functionals Castillo & Nickl

- **Setup : wavelet expansion** $f = \sum_j \sum_k f_{j,k}\phi_{j,k}$
  - Full BvM : posterior

$$(\sqrt{n}(f_{j,k} - \hat{f}_{j,k}), j \geq 1, k \in I_j) \to \mathcal{N}(0, K(.,.))$$

infinite dimensional vector : almost impossible —> weaker

# Weak BvM for smooth functionals Castillo & Nickl

- **Setup : wavelet expansion** $f = \sum_j \sum_k f_{j,k}\phi_{j,k}$
  - Full BvM : posterior

$$(\sqrt{n}(f_{j,k} - \hat{f}_{j,k}), j \geq 1, k \in I_j) \to \mathcal{N}(0, K(.,.))$$

infinite dimensional vector : almost impossible —> weaker

  - Weaker BvM : For all $J$, if $w_j \to +\infty$

$$(\sqrt{n}(f_{j,k} - \hat{f}_{j,k}), j \leq J, k \in I_j) \to \mathcal{N}(0, K_J(.,.))$$

$$E\left[\sup_{j,k} \frac{1}{w_j\sqrt{j}}|f_{j,k} - \hat{f}_{j,k}| \middle| X^n\right] = o_p(1/\sqrt{n})$$

$$\Rightarrow \text{ BvM for } \quad \psi(f) = \int \psi f dx, \sum_j w_j\sqrt{j}\sum_k |\psi_{j,k}| < +\infty$$

smooth functionals

# Conclusion and perspectives 1

- **Coverage of bias** Difficulties
  - Requires upper bound of $\pi(d(\theta, \theta_0) \leq \rho_n \epsilon_n)$

# Conclusion and perspectives 1

- **Coverage of bias** Difficulties
  - Requires upper bound of $\pi(d(\theta, \theta_0) \leq \rho_n \epsilon_n)$
  - Result for $C_n(L_n)) = \{d(\theta, \hat{\theta}) \leq L_n r_n(\alpha)\}$ Can we get rid of $L_n$ ? replace with $\alpha_n \to 1$ ?

# conclusion 2

- **Importance of the loss : semi-parametric** Model $\theta = (\psi, \eta)$ , $\psi \in \mathbb{R}^d$ .
  • Importance of a careful modelling of $\eta$ otherwise swamps everything $\rightarrow$ bad inference on $\psi$.
  • Robust inference ? change of likelihood
  - $$f_\theta(\mathbf{x}^n)^\tau, \quad \tau \in (0, 1)$$
- **Open questions : coverage and BvM**
  • What happens outside models equivalent to white noise ?
  • What can we say with more complex geometries : Mixture models ?

▶ **Importance of the loss : semi-parametric** Model $\theta = (\psi, \eta)$ , $\psi \in \mathbb{R}^d$ .

• Importance of a careful modelling of $\eta$ otherwise swamps everything $\rightarrow$ bad inference on $\psi$.

• Robust inference ? change of likelihood

- $$f_\theta(\mathbf{x}^n)^\tau, \quad \tau \in (0, 1)$$
- PAC Bayesian , Gibbs- type likelihood , Holmes and Walker

▶ **Open questions : coverage and BvM**

• What happens outside models equivalent to white noise ?

• What can we say with more complex geometries : Mixture models ?

# conclusion 2

- **Importance of the loss : semi-parametric** Model $\theta = (\psi, \eta)$ , $\psi \in \mathbb{R}^d$ .

• Importance of a careful modelling of $\eta$ otherwise swamps everything $\rightarrow$ bad inference on $\psi$.
• Robust inference ? change of likelihood

  - $$f_\theta(\mathbf{x}^n)^\tau, \quad \tau \in (0,1)$$
  - PAC Bayesian , Gibbs- type likelihood , Holmes and Walker
  - Bayes empirical likelihood

- **Open questions : coverage and BvM**

• What happens outside models equivalent to white noise ?
• What can we say with more complex geometries : Mixture models ?