
Discussion of talks by Judith Rousseau and Xinyi Xu

Iain Johnstone, Statistics, Stanford

`imj@stanford.edu`

OBayes 09, June, 2009

A 'Great Theorems' Session!

A 'Great Theorems' Session!

Bernstein-von Mises: (conditions under which) credible and confidence intervals agree to first order.

Pinsker: Beautifully packages ways for a linear estimate (in Gaussian data) to be optimal among non-linear estimators:

- ▲ θ very small
- ▲ θ very large
- ▲ θ such that a Gaussian prior is good

Today's results

Rousseau: A **semi-parametric** Bernstein-von Mises theorem

- ▲ impressive formulation, proof
- ▲ intriguing suggestion of counterexample

Xu: A **predictive** Pinsker theorem

- ▲ impressive formulation, proof
- ▲ explicit description of optimal estimators

Predictive density estimation

Parallels

Quadratic:

$$X \sim N_n(\theta, v_x I_n)$$

$$R_Q(\mu, \hat{\mu}) = E_\mu \|\hat{\mu} - \mu\|^2$$

$$\hat{\mu}_{MLE}(x) = x$$

- **formal Bayes for π_U**
- **minimax**
- **admissible for $p = 1, 2$**

Predictive:

$$Y \sim N_n(\theta, v_y I_n)$$

$$R_{KL}(\mu, \hat{\mu}) = \int p(x|\mu)p(y|\mu) \log \frac{p(y|\mu)}{p(x|\mu)} dx dy$$

$$\hat{p}_U(y|x)$$

- **formal Bayes for π_U**
- **minimax**
- **admissible for $p = 1$**

Parallels, II

Quadratic: Marginal density

$$\hat{\mu}_\pi = x + \nabla \log m_\pi(x)$$

[Brown, 1971]

Predictive:

$$\hat{p}_\pi(y|x) = \hat{p}_U(y|x) \mathbf{exp} \left(\log m_\pi(w, v_w) - \log m_\pi(x, v_x) \right),$$

[George, Liang, Xu, 2006]

$$\left[w = \frac{v_y x + v_x y}{v_x + v_y}, \quad v_w = \frac{v_x v_y}{v_x + v_y} \right]$$

Parallels, III

Quadratic: Stein's unbiased estimate of risk:

$$R_Q(\mu, \hat{\mu}_{MLE}) - R_Q(\mu, \hat{\mu}_\pi) = -2 \frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v) \Big|_{v=1}$$

[Brown, 1971]

Predictive:

$$R_{KL}(\mu, \hat{\mu}_U) - R_{KL}(\mu, \hat{p}_\pi) = \int_{v_x}^{v_w} \underbrace{\frac{\partial}{\partial v} E_{\mu, v} \log m_\pi(Z; v)}_{2E_{\mu, v} \left[\frac{\nabla^2 \sqrt{m_p(Z, v)}}{\sqrt{m_p(Z, v)}} \right]} dv$$

[George, Liang, Xu, 2006]

Pinsker's 'Little' Theorem

Quadratic: $\hat{\theta}_S = Sx/(S+1)$ is Bayes for $\pi_S = N(0, S)$

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R_Q(\theta, \hat{\theta}_S) = \sup_{\theta \in \Theta(C)} \inf_S R_Q(\theta, \hat{\theta}_S)$$

Pinsker, 1980

Predictive: \hat{p}_S linear predictive density for $\pi_S = N(0, S)$:

$$R_L(\Theta) = \inf_S \sup_{\theta \in \Theta(C)} R_{KL}(\theta, \hat{p}_S) = \sup_{\theta \in \Theta(C)} \inf_S R_{KL}(\theta, \hat{p}_S)$$

Xu-Liang, 2008

A Comment

Kneser-Kuhn minimax theorem: If C, D convex, D compact, and $(c, d) \rightarrow f(c, d)$ is convex-concave, and u.s.c. in d , then

$$\inf_c \sup_d f(c, d) = \sup_d \inf_c f(c, d).$$

Quadratic:

$$f(c, d) = \sum \epsilon^2 (1 - c_i)^2 + c_i^2 d_i \quad \text{is convex in } c_i = \frac{1}{1 + s}, \text{ linear in } d_i.$$

Predictive:

$$R(\theta, \hat{p}_S) \text{ is convex in } c_i = \frac{1}{v_{n+m} + s_i}, \text{ linear in } d_i = \theta_i^2.$$

Pinsker's "Big" theorem

Quadratic:

Linear minimax prior $N(0, S^*)$ "nearly" concentrates on $\Theta(C)$, so linear estimator "nearly" optimal.

Belister-Levit 1995,1996

Predictive:

Linear minimax prior $N(0, S^*)$ "nearly" concentrates on $\Theta(C)$, so linear predictive estimator \hat{p}_{S^*} is "nearly" optimal.

PLUS: predictive density complete class theorem, etc.

Bernstein-von Mises

Growing Gaussian Location (GGL) Model

Data: $\bar{Y} | \theta \sim N_p(\theta, \sigma_n^2 I)$ **Prior:** $\theta \sim N_p(0, \tau_n^2 I)$

[$p = p(n) \nearrow$, but **Gaussian** \Rightarrow **centering and scaling is key.**]

Posterior: Let $w_n = \tau_n^2 / (\sigma_n^2 + \tau_n^2)$. Then

$$\theta | \bar{Y} = \bar{y} \sim N_p(\hat{\theta}_B = w_n \bar{y}, w_n \sigma_n^2 I)$$

3 PERSPECTIVES:

1. **Global convergence of entire posterior**

2. **A non-linear functional:** $\|\theta - \hat{\theta}\|^2$

3. **Linear functionals** Lf (in $dY = f + \sigma_n dW$)

– progressively less ‘demanding’ for BvM

1. Global Convergence

When does

$$\|P_{\theta|Y^n} - N(\hat{\theta}_{MLE}, \frac{1}{nI_{\theta}})\| \xrightarrow{P_{\theta_0}^n} 0?$$

Here

$$\begin{aligned} P_{\theta|Y^n} &\leftrightarrow N_p(w_n \bar{y}, w_n \sigma_n^2 I) \\ N(\hat{\theta}_{MLE}, \frac{1}{nI_{\theta}}) &\leftrightarrow N_p(\bar{y}, \sigma_n^2 I) \end{aligned}$$

PROPOSITION **(BvM)** holds if and only if

$$\begin{aligned} \sqrt{p} \frac{\sigma_n^2}{\tau_n^2} &= \frac{\sqrt{p} \sigma_0^2}{n \tau_n^2} \rightarrow 0 \\ \text{i.e. } w_n &= 1 - o\left(\frac{1}{\sqrt{p_n}}\right) \end{aligned}$$

– a rate constraint on $w_n \rightarrow 1$.

2. A non-linear functional: $\|\hat{\theta}_B - \theta\|^2$

$$\theta|Y^n \sim N_p(\hat{\theta}_B = w_n \bar{y}, w_n \sigma_n^2 I) \quad w_n = \tau_n^2 / (\sigma_n^2 + \tau_n^2)$$

For $T_n = \|\theta - \hat{\theta}_B\|^2$, after **Freedman(1999)**, set $C_n = p\sigma^2 w_n$,

▲ for Bayesian: $T_n = C_n + \sqrt{D_n} Z_{1n}$

▲ for Frequentist: $T_n = C_n + \sqrt{F_n} Z_{2n}(\theta) + \sqrt{G_n(\theta)} Z_{3n}(\theta, \epsilon)$

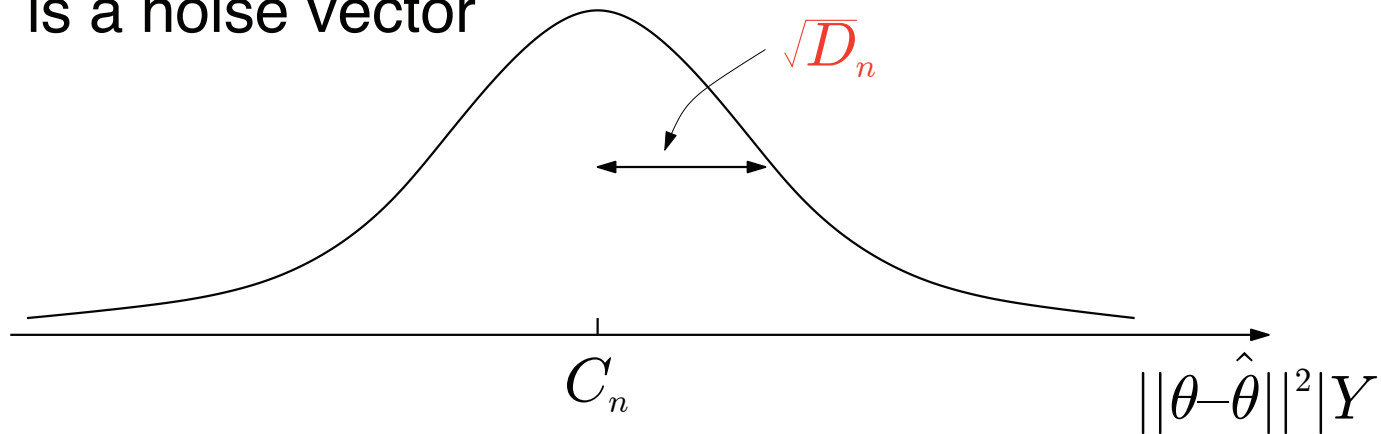
$$\sqrt{D_n} = \text{Bayesian SD} = \sqrt{2p\sigma^2} \cdot w_n$$

$$\sqrt{G_n(\theta)} = \text{Frequentist SD} = \sqrt{2p\sigma^2} \cdot w_n [2w_n - w_n^2]^{1/2}$$

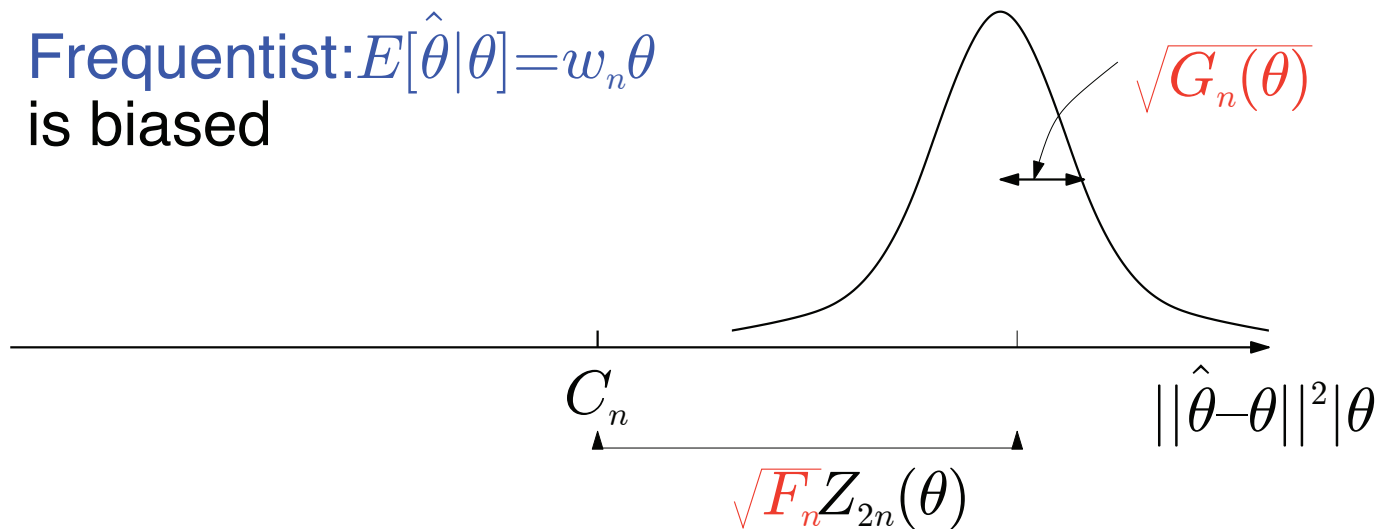
$$\sqrt{F_n} = \text{'wobble' in Freq. mean} = \sqrt{2p\sigma^2} \cdot w_n (1 - w_n)$$

2. BvM fails for $\|\theta - \hat{\theta}\|^2$ if $\lim w_n < 1$

Bayesian: $\theta - \hat{\theta}$
is a noise vector



Frequentist: $E[\hat{\theta} | \theta] = w_n \theta$
is biased



If $\lim w_n = 1$, $\sqrt{F_n} = o(\sqrt{D_n})$, **Bayes SD \sim Freq. SD.**

3. Linear Functionals Lf

For any o.n. basis $\{\psi_\lambda(t)\}$, $Y_t = \int_0^t f(s)ds + \sigma_n W_t \quad t \in [0, 1]$

$$\Leftrightarrow y_\lambda = \theta_\lambda + \sigma_n z_\lambda \quad z_\lambda \sim N(0, 1).$$

$$f(t) = \sum \theta_\lambda \psi_\lambda(t) \quad \Rightarrow \quad Lf = \sum \theta_\lambda L\psi_\lambda = \sum \theta_\lambda a_\lambda, \quad \text{say}$$

Prior: $\theta_\lambda \stackrel{\text{ind}}{\sim} N(0, \tau_\lambda^2)$

Posterior: $\theta_\lambda | y_\lambda \stackrel{\text{ind}}{\sim} N(w_\lambda y_\lambda, w_\lambda \sigma_n^2), \quad w_\lambda = \tau_\lambda^2 / (\sigma_n^2 + \tau_\lambda^2)$

Bayesian: $Lf | y \sim N(L\hat{f} = \sum a_\lambda w_\lambda y_\lambda, V_y = \sigma_n^2 \sum a_\lambda^2 w_\lambda)$

Frequentist: $\widehat{L}f | f \sim N(E\widehat{L}f, V_F = \sigma_n^2 \sum a_\lambda^2 w_\lambda^2)$

3. Bounded Functionals

Variance Ratio: $\frac{V_F}{V_y} = \frac{\sum a_\lambda^2 w_\lambda^2}{\sum a_\lambda^2 w_\lambda} < 1, \quad w_\lambda \rightarrow 1 \quad (\text{fixed } \lambda!)$

a) Bounded linear functionals: $\sum_\lambda a_\lambda^2 < \infty \Leftrightarrow \int a^2 < \infty.$

e.g. $a(t) = \sum_0^K c_k t^k$, **polynomials**
 $= I_B(t)$ **region of interest, c.d.f.**

$V_F/V_y \rightarrow 1$, & **Bayesians & frequentists use same intervals.**

b) (Badly) unbounded functionals: have $\lim V_F/V_y < 1.$

e.g. values at a point: $Lf = f^{(r)}(t_0)$: if $\tau_\lambda^2 \sim 2^{-2jm}$, then

$$\frac{V_F}{V_y} \rightarrow 1 - \frac{2r + 1}{2m} \Rightarrow \text{BvM fails.}$$