Confidence regions in high-dimensional and nonparametric statistical models

Richard Nickl

Statistical Laboratory, DPMMS
University of Cambridge (UK)

Eindhoven, Jan 2017, YES VIII

Introduction

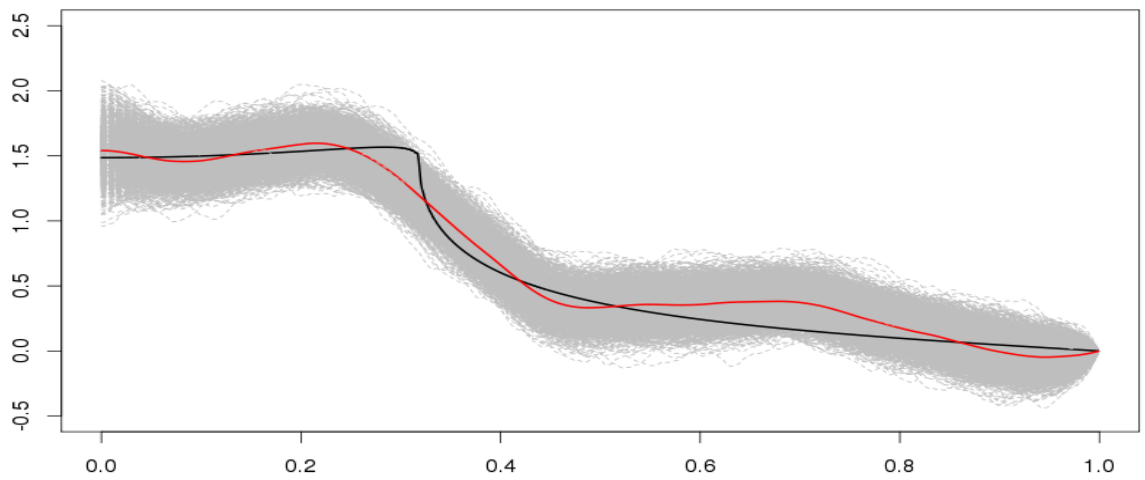
Given observations Y from some model $\{P_\theta : \theta \in \Theta\}$ in mind, statistical inference consists of three fundamental tasks:

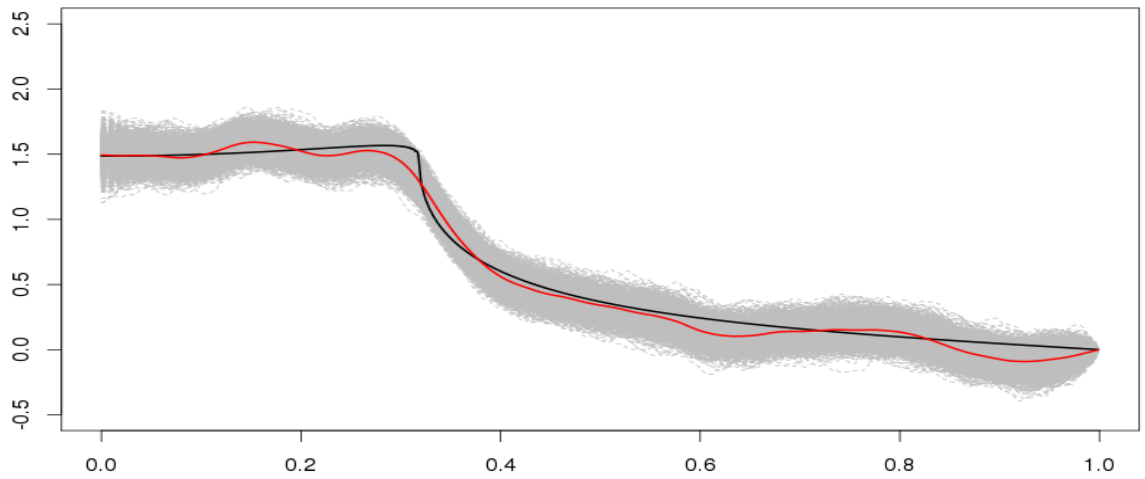
1) Parameter Estimation: the construction of an algorithm $T(Y)$ that estimates the unknown θ, hopefully 'optimally' in some sense.

2) Hypothesis testing: given a hypothesis H_0 on the parameter θ, the construction of a test function $\Psi(Y)$ that indicates whether H_0 is likely to be true or not.

3) Uncertainty quantification: construction of a 'smallest possible' confidence region $C(Y) \subset \Theta$ such that $P_\theta(\theta \in C(Y))$ with high probability.

Point estimate in red of f at sample sizes $n = 500, 2000,$ & confidence intervals.





Traditional parametric inference

For example when using the maximum likelihood estimator $\hat{\theta}_{MLE}$ in a regular parametric model $\{P_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^p$, one has as $n \to \infty$ (and p fixed),

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) \to^d N(0, I(\theta_0)^{-1}) \quad \text{under } P_{\theta_0},\ \theta_0 \in \Theta;$$

and using a consistent estimate $\hat{i}_n$ of the Fisher information $I(\theta_0)$ and some quantile constants z_α, one obtains an asymptotic confidence interval ($p = 1$)

$$C_n = \left[\hat{\theta}_{MLE} - \frac{z_\alpha \hat{i}_n}{\sqrt{n}}, \hat{\theta}_{MLE} + \frac{z_\alpha \hat{i}_n}{\sqrt{n}}\right].$$

$\to$ These quantile constants can also be replaced by:

i) bootstrap estimates, or by
ii) quantiles of suitable Bayesian posterior distributions.

$\to$ The latter practice can be justified rigorously by *bootstrap consistency and Bernstein-von Mises theorems*.

Nonparametric statistical models

Nonparametric density estimation: $Y_1, \ldots, Y_n$ i.i.d. from unknown probability density θ such that its Sobolev norm satisfies

$$\|\theta\|_{H^\beta} = \sum_{0 \le \alpha \le \beta} \|D^\alpha \theta\|_{L^2} < \infty, \quad \beta \in \mathbb{N}.$$

Nonparametric regression: For response and design variables (Y_i, X_i),

$$Y_i = \theta(X_i) + \varepsilon_i, \quad i = 1, \ldots, n; \ \varepsilon_i \sim^{i.i.d.} N(0, \sigma^2)$$

where θ, $\|\theta\|_{H^\beta} < \infty$, is an unknown regression function, σ^2 the noise variance.

$\rightarrow$ Minimax convergence rates here are of the form

$$\inf_{\hat{\theta}} \sup_{\theta : \|\theta\|_{H^\beta} \le B} E_\theta \|\hat{\theta} - \theta\|_{L^2}^2 \simeq n^{-2\beta/(2\beta+1)} \gg \frac{1}{n}.$$

$\rightarrow$ One can heuristically think of these models as such that the model dimension

$$p \approx n^{1/(2\beta+1)} \ \text{ as } n \to \infty.$$

The asymptotic regime is fundamentally different, particularly if β is unknown.

High-dimensional models I

Linear model: For 'regular' design variables $X = (X_{ij})$, let

$$Y_i = \sum_{j=1}^{p} X_{ij}\theta_j + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} N(0, \sigma^2), \ i = 1, \dots, n.$$

We think here of $p = p_n \to \infty$ possibly with $p \geq n$, and θ is k-sparse, where $k = k_n$ may also depend on n. The set of k-sparse vectors is denoted by

$$B_0(k) = \{\theta \in \mathbb{R}^p, \theta_j \neq 0 \text{ for at most } k \text{ indices } j\}$$

$\to$ For $k = o(n)$, the minimax performance in squared loss is

$$\inf_{\hat{\theta}} \sup_{\theta \in B_0(k)} E_\theta \|\hat{\theta} - \theta\|^2 \simeq \frac{k \log p}{n}.$$

$\to$ The above bound is attained by computationally efficient estimators such as the Lasso $\hat{\theta}$ which minimises

$$\|Y - X\theta\|^2_{\mathbb{R}^n} + \lambda \sum_{j=1}^{p} |\theta_j| \quad \text{over } \mathbb{R}^p.$$

High-dimensional models II

$\rightarrow$ A similar theory exists for matrix recovery/completion problems

$$Y_i = tr(X_i\theta) + \varepsilon_i, \;\; i = 1, \dots, n$$

where sparsity is replaced by low rank conditions on the matrix θ.

$\rightarrow$ Let $\theta \in R_{m_1 m_2}(k)$ mean that θ is a $m_1 \times m_2$ matrix of rank k. Then under suitable conditions, the minimax performance is

$$\inf_{\hat{\theta}} \sup_{\theta \in R_{m_1 m_2}(k)} E_\theta \|\hat{\theta} - \theta\|_F^2 \simeq \frac{k(m_1 + m_2)}{n}.$$

$\rightarrow$ The sensing matrices X_i may satisfy a 'restricted isometry property', or not: in matrix completion $trace(X_i\theta)$ reports a randomly picked matrix entry of θ.

$\rightarrow$ The matrix Lasso $\hat{\theta}$ can be used; for $\mathcal{X}\theta = (tr(X_1\theta) \dots tr(X_n\theta))^T$, it minimises

$$\|Y - \mathcal{X}\theta\|_{\mathbb{R}^n}^2 + \lambda \|\theta\|_{S_1} \;\; \text{over } M_{m_1, m_2}.$$

Uniform inference with adaptive estimators?

- Can we construct a confidence set, that is, choose $R_n(\alpha)$ in

$$C_n = \left\{\theta : \|\theta - \hat{\theta}\| \leq R_n(\alpha)\right\}$$

such that for some 'optimal' rate $r_{n,opt}$ we have

$$\inf_{\theta_0} P_{\theta_0}(\theta_0 \in C_n) \geq 1 - \alpha - o(1), \quad R_n(\alpha) = O_p(r_{n,opt})?$$

- Penalisation estimators as the Lasso cannot be obviously used for inference, initially just because their distributions are not easily obtained, and in fact cannot reliably be estimated either.

- There is no alternative: If $p \geq n$, some dimension reduction is necessary, as 'classical' estimators in the 'maximal' model are not sensible any longer.

- We want confidence regions for the **full parameter** $\theta \in \mathbb{R}^p$, and not just for one fixed entry θ_1 or another simple linear functional of θ.

A basic minimax framework for adaptive confidence regions

- Consider observations $(Y_n \sim P_\theta^n : n \in \mathbb{N}), \theta \in \Theta$, where Θ is equipped with some metric d. Suppose that

over Θ the minimax rate of estimation is $r_n(\Theta)$.

A basic minimax framework for adaptive confidence regions

- Consider observations $(Y_n \sim P_\theta^n : n \in \mathbb{N}), \theta \in \Theta$, where Θ is equipped with some metric d. Suppose that

  over Θ the minimax rate of estimation is $r_n(\Theta)$.

- Then consider a sub-model ('the adaptation/selection hypothesis')

  $\Theta_0 \subset \Theta$ where the minimax rate of estimation is $r_n(\Theta_0) = o(r_n(\Theta))$.

A basic minimax framework for adaptive confidence regions

- Consider observations $(Y_n \sim P_\theta^n : n \in \mathbb{N}), \theta \in \Theta$, where Θ is equipped with some metric d. Suppose that

  over Θ the minimax rate of estimation is $r_n(\Theta)$.

- Then consider a sub-model ('the adaptation/selection hypothesis')

  $\Theta_0 \subset \Theta$ where the minimax rate of estimation is $r_n(\Theta_0) = o(r_n(\Theta))$.

- Ex 1: Nonparametric function estimation

$$r_n^2(\Theta) \approx n^{-2\beta/(2\beta+1)}, \quad r_n^2(\Theta_0) \approx n^{-2\beta_0/(2\beta_0+1)}, \quad \beta_0 > \beta.$$

A basic minimax framework for adaptive confidence regions

- Consider observations $(Y_n \sim P_\theta^n : n \in \mathbb{N}), \theta \in \Theta$, where Θ is equipped with some metric d. Suppose that

  over Θ the minimax rate of estimation is $r_n(\Theta)$.

- Then consider a sub-model ('the adaptation/selection hypothesis')

  $\Theta_0 \subset \Theta$ where the minimax rate of estimation is $r_n(\Theta_0) = o(r_n(\Theta))$.

- Ex 1: Nonparametric function estimation

$$r_n^2(\Theta) \approx n^{-2\beta/(2\beta+1)}, \quad r_n^2(\Theta_0) \approx n^{-2\beta_0/(2\beta_0+1)}, \quad \beta_0 > \beta.$$

- Ex 2: High-dimensional sparse regression

$$r_n^2(\Theta) \approx \frac{k \log p}{n}, \quad r_n^2(\Theta_0) \approx \frac{k_0 \log p}{n}, \quad k_0 = o(k).$$

A basic minimax framework for adaptive confidence regions

- Consider observations $(Y_n \sim P_\theta^n : n \in \mathbb{N}), \theta \in \Theta$, where Θ is equipped with some metric d. Suppose that

  over Θ the minimax rate of estimation is $r_n(\Theta)$.

- Then consider a sub-model ('the adaptation/selection hypothesis')

  $\Theta_0 \subset \Theta$ where the minimax rate of estimation is $r_n(\Theta_0) = o(r_n(\Theta))$.

- Ex 1: Nonparametric function estimation

$$r_n^2(\Theta) \approx n^{-2\beta/(2\beta+1)}, \quad r_n^2(\Theta_0) \approx n^{-2\beta_0/(2\beta_0+1)}, \quad \beta_0 > \beta.$$

- Ex 2: High-dimensional sparse regression

$$r_n^2(\Theta) \approx \frac{k \log p}{n}, \quad r_n^2(\Theta_0) \approx \frac{k_0 \log p}{n}, \quad k_0 = o(k).$$

- Ex 3: Low rank matrix estimation

$$r_n^2(\Theta) \approx \frac{k(m_1 + m_2)}{n}, \quad r_n^2(\Theta_0) \approx \frac{k_0(m_1 + m_2)}{n}, \quad r_0 = o(r).$$

Honest adaptive confidence regions

- We want a confidence set C_n with the following properties: If

$$|C|_d = \sup\{d(x,y) : x, y \in C\}$$

is the d-diameter of a set $C \subset \Theta$, then for n large enough

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n) \geq 1 - \alpha, \qquad \textit{"honest"}$$

$$\sup_{\theta \in \Theta_0} E_\theta |C_n|_d \lesssim r_n(\Theta_0), \qquad \textit{"adaptive" to sub-model}$$

$$\sup_{\theta \in \Theta} E_\theta |C_n|_d \lesssim r_n(\Theta).$$

For 'negative' results we may use 'in probability' versions of the last two requirements.

Minimax testing rates I

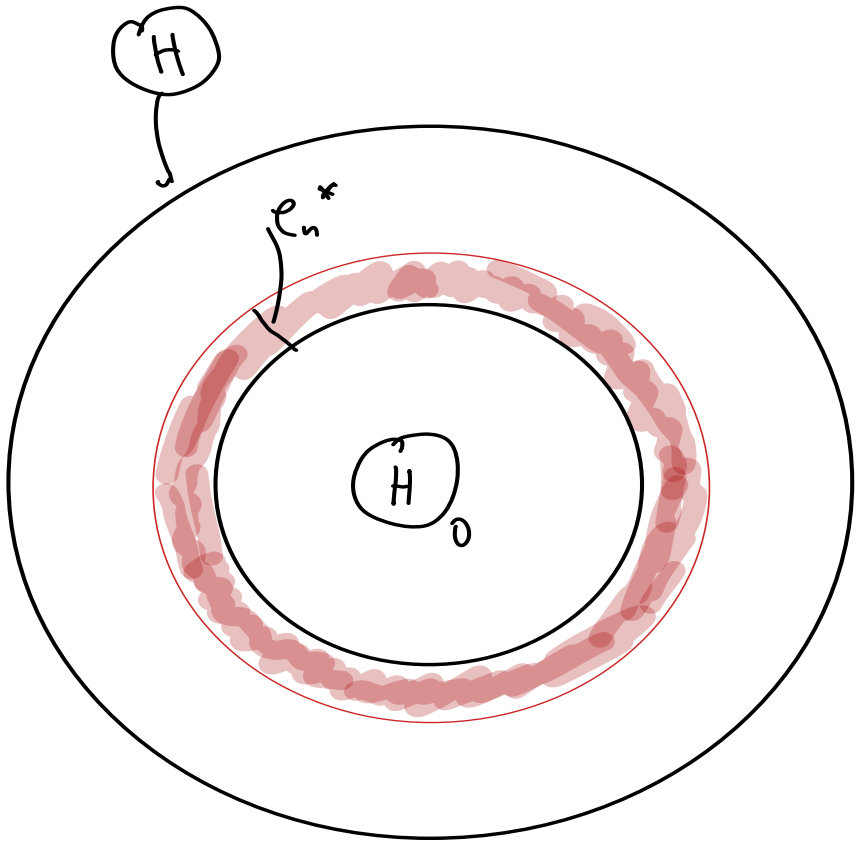
- We introduce the minimax testing rate for the hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 = H_1(\rho_n) = \{\theta \in \Theta, d(\theta, H_0) \geq \rho_n\}, \ \rho_n \geq 0.$$

Minimax testing rates I

- We introduce the minimax testing rate for the hypothesis testing problem

$$H_0 : \theta \in \Theta_0 \quad vs \quad H_1 = H_1(\rho_n) = \{\theta \in \Theta, d(\theta, H_0) \geq \rho_n\}, \ \rho_n \geq 0.$$



Minimax testing rates II

- $H_0 : \theta \in \Theta_0 \quad vs \quad H_1 = H_1(\rho_n) = \{\theta \in \Theta, d(\theta, H_0) \geq \rho_n\}, \ \rho_n \geq 0.$

Minimax testing rates II

- $H_0 : \theta \in \Theta_0 \quad vs \quad H_1 = H_1(\rho_n) = \{\theta \in \Theta, d(\theta, H_0) \geq \rho_n\}, \ \rho_n \geq 0.$
- In words: ρ_n^* is the minimal sequence ρ_n required to be able to distinguish the hypotheses H_0 and $H_1(\rho_n)$ consistently by a test function $\Psi(Y_n) \in \{0, 1\}$.

Minimax testing rates II

- $H_0 : \theta \in \Theta_0 \quad vs \quad H_1 = H_1(\rho_n) = \{\theta \in \Theta, d(\theta, H_0) \geq \rho_n\}, \ \rho_n \geq 0.$
- In words: ρ_n^* is the minimal sequence ρ_n required to be able to distinguish the hypotheses H_0 and $H_1(\rho_n)$ consistently by a test function $\Psi(Y_n) \in \{0, 1\}$.

Definition of the minimax testing rate $(\rho_n^* : n \in \mathbb{N})$

i) for $\rho_n \geq \rho_n^*$ and every $\alpha > 0$ there exists a test function $\Psi_n = \Psi(Y_n, \alpha)$ s.t.

$$\sup_{\theta \in H_0} P_\theta(\Psi_n = 1) + \sup_{\theta \in H_1(\rho_n)} P_\Theta(\Psi_n = 0) \leq \alpha,$$

ii) for any sequence $\rho_n = o(\rho_n^*)$ we have that

$$\inf_{\Psi_n} \left[\sup_{\theta \in H_0} P_\theta(\Psi_n = 1) + \sup_{\theta \in H_1(\rho_n)} P_\theta(\Psi_n = 0) \right] \geq \beta_0.$$

- In most situations below we will in fact have $\beta_0 = 1$.

Heuristic information geometry

Hypothesis testing and confidence sets

The two theorems that follow show that for a parameter space Θ endowed with metric d, the pair

$$\rho_n^*, \quad r_n(\Theta_0)$$

corresponding to the 'model selection' problem

$$\Theta_0 \subset \Theta$$

entirely determines the answer to the question of whether adaptive confidence sets (with respect to the metric d) exist.

Testing rate vs adaptive estimation rate over Θ_0

Effectively one has to check which of these rates approaches zero faster, that is,

$$\text{whether } r_n(\Theta_0) = o(\rho_n^*) \text{ or } \rho_n^* = o(r_n(\Theta_0)).$$

A lower bound

Theorem

Let ρ_n^* be the minimax rate of testing, and assume

$$r_n(\Theta_0) = o(\rho_n^*)$$

as $n \to \infty$. Then *any* confidence set C_n that is honest over Θ, i.e., such that

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta} P_\theta(\theta \in C_n) \geq 1 - \alpha$$

for some $\alpha < \beta_0/3$, necessarily satisfies for some constant $c = c(\alpha)$,

$$\sup_{\theta\in\Theta_0} E_\theta |C_n|_d \geq c\rho_n^* \gg r_n(\Theta_0).$$

• Hence lower bounds on the testing rate ρ_n^* can be used to prove non-existence of confidence regions if the rate $r_n(\Theta_0)$ of estimation in the 'adaptation hypothesis' Θ_0 is of smaller order than that lower bound.

A converse upper bound

• Let us say that an estimator $\hat{\theta}$ satisfies an oracle inequality at level $\beta > 0$ if for some $C > 0$ and all $\theta \in \Theta$ we have with $\mathbb{P}_\theta^n$-probability $\geq 1 - \beta$,

$$d(\hat{\theta}, \theta) \leq C \inf_{T \in \{\Theta, \Theta_0\}} (d(\theta, T) + r_n(T)).$$

Theorem

Let ρ_n^* be the minimax rate of testing, assume as $n \to \infty$,

$$\rho_n^* = o(r_n(\Theta_0)),$$

and suppose an 'oracle' estimator $\hat{\theta}$ exists. Then there exists a confidence set $C_n = C_n(Y, \alpha)$ that is honest over Θ, i.e.,

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

for every $\alpha > 0$ and n large enough, and such that for all $n \in \mathbb{N}$,

$$\sup_{\theta \in \Theta_0} E_\theta |C_n|_d \lesssim r_n(\Theta_0), \quad \sup_{\theta \in \Theta} E_\theta |C_n|_d \lesssim r_n(\Theta).$$

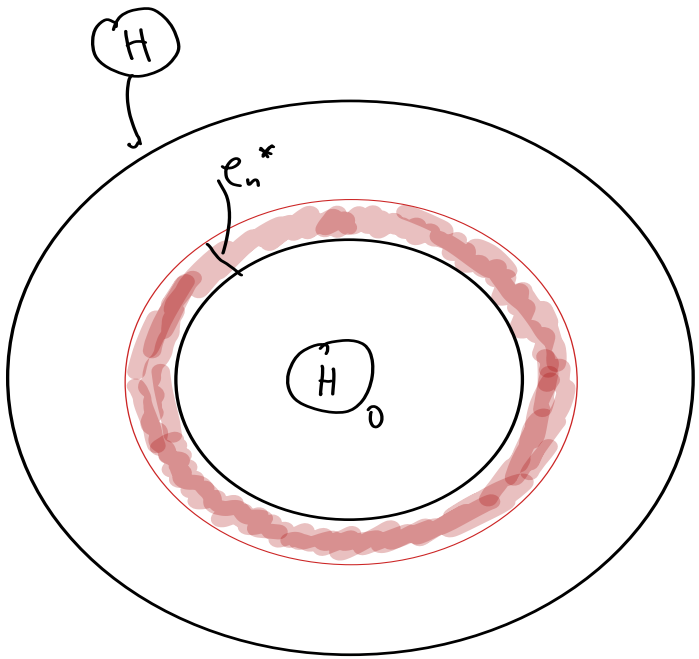
Signal strength conditions

• When $r_n(\Theta_0) = o(\rho_n^*)$ the proofs of the above theorems in fact imply that adaptive confidence sets do exist if and only if we restrict the parameter space to

$$\Theta_n^* = \Theta_0 \cup \{\theta \in \Theta : d(\theta, \Theta_0) \geq \rho_n^*\} \subseteq \Theta, \quad \rho_n^* \geq 0,$$

so that we have removed critical 'untestable signals' too close to Θ_0;

$$\{\theta \in \Theta : d(\theta, \Theta_0) < \rho_n^*\}.$$



So what?

• Before we discuss the proofs, and move to some examples, let us point out why there is some hope to obtain new insights.

• In parametric models the testing and estimation rates are usually all of order $1/\sqrt{n}$, so the above theory is not necessarily informative.

• But in high - or infinite dimensional models the testing and estimation rates can be of quite different orders of magnitude. This has been pointed out in in seminal work by Yuri Ingster and others in a sequence of papers in the 1990s and 2000s, where the signal detection problem is studied

$$H_0 : f = 0 \quad vs. \; H_1 = \{f \in H^\beta, \|f\|_{L^2} \geq \rho\},$$

and where it is shown that

$$\rho_n^* \approx n^{-\beta/(2\beta+1/2)} = o(n^{-\beta/(2\beta+1)}).$$

• Our 'composite' testing problem is more involved but some intuitions and ideas carry over from Ingster's work.

Some proof ideas I: lower bounds

• We argue by contradiction, and suppose an adaptive confidence set C_n exists of honest coverage $1-\alpha$. Since $r_n(\Theta_0) = o(\rho_n^*)$ we can choose a sequence $\bar{\rho}_n$ such that

$$r_n(\Theta_0) \ll \bar{\rho}_n \ll \rho_n^*$$

• We can test

$$H_0 : \theta \in \Theta_0, \quad H_1 = H_1(\rho_n) : \{\theta : d(\theta, \Theta_0) \geq \rho_n\}$$

by

$$\Psi_n = \mathbf{1}\{C_n \cap H_1 \neq \emptyset\},$$

so we reject H_0 if C_n contains any of the alternatives.

• For the type-two errors of this test we obviously have

$$P_\theta(\Psi_n = 0) \leq P_\theta(C_n \cap H_1 = \emptyset, \theta \in C_n) + P(\theta \notin C_n) \leq 0 + \alpha$$

since $\theta \in H_1$ implies that $C_n \cap H_1$ is non-empty in the middle event.

Some proof ideas II: lower bounds, ct'd

- For the type-one errors when $\theta \in \Theta_0$, we have

$$\begin{aligned} P_\theta(\Psi_n = 1) &= P_\theta(C_n \cap H_1 \neq \emptyset) \\ &\leq P_\theta(C_n \cap H_1 \neq \emptyset, \theta \in C_n) + P_\theta(\theta \notin C_n) \\ &\leq P_\theta(|C_n|_d \geq \bar{\rho}_n) + \alpha \leq 2\alpha \end{aligned}$$

since adaptivity of C_n implies that $|C_n|_d = O(r_n(\Theta_0)) = o(\bar{\rho}_n)$.

- We conclude that this test has error level

$$\sup_{\theta \in H_0} P_\theta(\Psi_n = 1) + \sup_{\theta \in H_1(\rho_n)} P_\Theta(\Psi_n = 0) \leq 3\alpha,$$

but since $\bar{\rho}_n = o(\rho_n^*)$ this contradicts, for $\alpha < \beta_0/3$ small enough, the lower bound

$$\inf_{\Psi_n} \left[\sup_{\theta \in H_0} P_\theta(\Psi_n = 1) + \sup_{\theta \in H_1(\rho_n)} P_\theta(\Psi_n = 0) \right] \geq \beta_0.$$

Some proof ideas III: upper bounds

• If the testing rate satisfies $\rho_n^* = O(r_n(\Theta_0))$, then there exists a consistent test Ψ_n for the hypotheses

$$H_0 : \theta \in \Theta_0, \quad H_1 = H_1(\rho_n^*) : \{\theta : d(\theta, \Theta_0) \geq \rho_n^*\}.$$

We then take as confidence set

$$C_n = \left\{\theta \in \Theta : d(\theta, \hat{\theta}) \leq L\left[r_n(\Theta_0)(1 - \Psi_n) + r_n(\Theta)\Psi_n\right]\right\},$$

where L is a large constant and $\hat{\theta}$ the oracle estimator.

• This confidence region will obviously work whenever the test works at suitably small level α. In the 'critical' region where

$$d(\theta, \Theta_0) \leq \rho_n^* = O(r_n(\Theta_0))$$

we still have, in view of the oracle inequality, that

$$d(\theta, \hat{\theta}) \leq d(\theta, \Theta_0) + r_n(\Theta_0) \lesssim r_n(\Theta_0),$$

so that C_n retains coverage.

Some applications of the general theory

Let us now see what the above approach yields in a few important examples:

- Nonparametric function estimation (regression)
- Sparse linear regression
- Low rank matrix recovery

Recall that we want a confidence set C_n with the following properties: If $|C|_d$ is the d-diameter of a set $C \subset \Theta$, then for n large enough

$$\inf_{\theta \in \Theta} P_\theta(\theta \in C_n) \geq 1 - \alpha, \qquad \text{"honest"}$$

$$\sup_{\theta \in \Theta_0} E_\theta |C_n|_d \lesssim r_n(\Theta_0), \qquad \text{"adaptive" to sub-model}$$

$$\sup_{\theta \in \Theta} E_\theta |C_n|_d \lesssim r_n(\Theta).$$

• To decide whether such C_n exists we need to obtain suitable upper and lower bounds on minimax testing rates in each statistical model. See Giné and Nickl (2016, Chapter VI.2) for some general principles.

Confidence sets for function estimation in L^∞-loss

Consider a density or regression function θ in a Hölder space $C^\beta([0,1]), \beta > 0$.

Theorem (Low, Hoffmann, Nickl)

1) A honest confidence set for $\Theta = \{\theta : \|\theta\|_{C^\beta} \leq B\}$ *that adapts to*

$$\Theta_0 = \{\theta : \|\theta\|_{C^{\beta_0}} \leq B\}, \beta_0 > \beta,$$

in L^∞*-loss does not exist . Any* $1-\alpha$ *honest confidence set* C_n *satisfies*

$$\sup_{\theta \in \Theta_0} E_\theta |C_n|_\infty \geq c_\alpha r_n(\Theta).$$

2) A honest adaptive confidence set over

$$\Theta_0 \cup \{\theta \in \Theta : \|\theta - \Theta_0\|_\infty \geq \rho_n\} \text{ exists if and only if } \rho_n \gtrsim \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+1}}$$

3) A sufficient condition is self-similarity: for K_j *a wavelet projector,*

$$\epsilon 2^{-j\beta} \leq \|K_j(f) - f\|_\infty \leq C 2^{-j\beta} \quad \forall j \geq J_0.$$

A limit theorem and 'genericity' of self-similarity

- Common adaptive estimators such as a density estimator $\hat{f}_n$ based on Lepski's method satisfy exact limit theorems under self-similarity conditions:

$$a_n(\|\hat{f}_n - f\|_\infty - b) \to^d Z, \quad Z \sim \text{Gumbel},$$

with known and optimal constants a_n, b_n, giving rise to adaptive confidence bands (Giné and Nickl (2010)).

A limit theorem and 'genericity' of self-similarity

- Common adaptive estimators such as a density estimator $\hat{f}_n$ based on Lepski's method satisfy exact limit theorems under self-similarity conditions:

$$a_n(\|\hat{f}_n - f\|_\infty - b) \to^d Z, \quad Z \sim \text{Gumbel},$$

with known and optimal constants a_n, b_n, giving rise to adaptive confidence bands (Giné and Nickl (2010)).

- There are several ways in which 'self-similarity' is a sensible model for β-Hölderian functions.

Theorem (Giné, Nickl)

The exceptional set $\mathcal{N}$ of functions in the Hölder space C^β that are NOT self-similar is nowhere dense in the norm topology of C^β.

- In other words: non-selfsimilar functions contain no Banach ball of C^β.
- Stefan Banach used this notion to prove that 'almost every continuous functions is nowhere differentiable'.

The Bayesian does not see non-selfsimilar functions

• Natural Bayesian priors Π_β on β-Hölder smooth functions satisfy

$$\Pi_\beta(f \in C^\beta \text{ is not self-similar}) = 0,$$

see Chapter VIII of Giné and Nickl (2016). From the perspective of 'Bayes-risk', the pathological functions for which honest inference is impossible do not exist.

• For instance, Gloter and Hoffmann (2007, AoS) show that if the unknown regression function is drawn from a fractional Brownian motion B_H with unknown Hurst index H, then the unknown smoothness of the function can be consistently estimated

$$|\hat{H} - H| = O_P(n^{-1/(4H+2)})$$

under the Bayesian model (so assuming $f \sim B_H$).

Confidence sets for function estimation in L^2-loss

Consider now density estimation or nonparametric regression with θ belonging to a Sobolev space $H^\beta([0,1])$, and L^2-loss.

Theorem (Robins, van der Vaart, Bull, Nickl, Szabo)

1) A honest confidence set for $\Theta = \{\theta : \|\theta\|_{H^\beta} \leq B\}$ *that adapts to*

$$\Theta_0 = \{\theta : \|\theta\|_{H^{\beta_0}} \leq B\}, \beta_0 > \beta,$$

in L^2*-loss exist if and only if* $\beta_0 \leq 2\beta$.

2) If $\beta_0 > 2\beta$, *a honest adaptive confidence set over*

$$\Theta_0 \cup \{\theta \in \Theta : \|\theta - \Theta_0\|_{L^2} \geq \rho_n\} \text{ exists if and only if } \rho_n \gtrsim n^{-\frac{\beta}{2\beta+1/2}}.$$

3) If $\beta_0 > 2\beta$, *a honest adaptive confidence region exist over* $\Theta_0 \cup \Theta_\rho$ *where*

$$\Theta_\rho \equiv \left\{ \theta : \sum_{k=N}^{N^\rho} \theta_k^2 \geq cN^{-2\beta} \quad \forall N \in \mathbb{N} \right\}, \quad \rho > 1; \ \theta_k = \langle \theta, e^{ik\cdot} \rangle_{L^2}.$$

Extensions

• The nonparametric theory is now fairly well understood, and important recent contributions include the paper by Szabo, van der Vaart and van Zanten (Ann.Stat. 2015), with several discussion contributions.

• The theory of nonparametric adaptive confidence sets introduced above (in the way I see it) is presented in more depth in Chapter 8.3 of the book Giné and Nickl (2016).

• Let us now move to 'high-dimensional models' where the theory is currently still in development.

Confidence sets in high-dimensional sparse regression

- In the model

$$Y_i = \sum_{j=1}^{p} X_{ij}\theta_j + \varepsilon_i, \quad \varepsilon_i \sim^{i.i.d.} N(0, \sigma^2), \ i = 1, \dots, n,$$

we assume that $\theta \in \Theta = B_0(k)$ is at least k-sparse for some $k = o(n)$, but we want a confidence set C_n that adapts to the sub-model $\Theta_0 = B_0(k_0)$ of k_0-sparse signals with $k_0 = o(k)$.

- More precisely, for n large enough,

$$\inf_{\theta \in B_0(k)} P_\theta(\theta \in C_n) \geq 1 - \alpha, \qquad \textit{"honest"}$$

$$\sup_{\theta \in B_0(k_0)} E_\theta |C_n|_2 \lesssim \sqrt{\frac{k_0 \log p}{n}}, \qquad \textit{"adaptive" to sub-model}$$

$$\sup_{\theta \in B_0(k)} E_\theta |C_n|_2 \lesssim \sqrt{\frac{k \log p}{n}}.$$

Confidence sets in high-dimensional sparse regression II

Theorem (Nickl, van de Geer)

Consider for simplicity i.i.d. Gaussian design (X_{ij}) *and* $\sigma^2 = 1$.

1) Assume $p \geq n, k \geq \sqrt{n}$. *A honest confidence set for* $\Theta = B_0(k)$ *that adapts to* $\Theta_0 = B_0(k_0)$ *exist if and only if*

$$r_n(\Theta_0) = \sqrt{\frac{k_0 \log p}{n}} \geq Cn^{-1/4} \quad C > 0 \text{ universal.}$$

2) In particular, and even if $p \leq n$, *no honest confidence confidence set exists for* Θ *that adapts to submodels* $\Theta_0 = B_0(k_0)$ *with*

$$k_0 = o(\sqrt{n/\log p}).$$

3) A honest adaptive confidence region exists over $\Theta_0 \cup \Theta_n$ *where*

$$\Theta_n \equiv \left\{ \theta \in \Theta \setminus \Theta_0 : \sum_{j=k_0+1}^{p} \theta_{(j)}^2 \geq Cn^{-1/2} \right\}, \quad \theta_{(j)}^2 \geq \theta_{(j+1)}^2 \quad \forall j$$

Confidence sets in high-dimensional sparse regression III

- When the noise variance σ^2 is known, one can take as confidence set

$$C_n = \left\{ \theta \in \mathbb{R}^p : \|\theta - \hat{\theta}\|^2 \leq \hat{R}_n + \frac{z_\alpha}{\sqrt{n}} \right\}$$

where z_α are some quantile constants and where

$$\hat{R}_n = \frac{1}{n}(Y - X\hat{\theta})^T(Y - X\hat{\theta}) - \sigma^2,$$

which gives a honest confidence set of diameter

$$\sup_{\theta \in B_0(k)} E_\theta |C|_2 \lesssim n^{-1/4} + \sqrt{\frac{k \log p}{n}}$$

which is adaptive as soon as $k \geq \sqrt{n/\log p}$ (see Nickl and van de Geer (2013)).

- Already when the noise variance is unknown this is not possible any more (Cai and Guo (2016)).

Confidence sets for low rank matrix recovery I

- Let us turn to matrix recovery/completion problems

$$Y_i = tr(X_i\theta) + \varepsilon_i, \ \ i = 1, \dots, n$$

with $m_1 \times m_2$ matrices θ in $\Theta = R_{m_1 m_2}(k)$ of rank at most k. The noise is independent with possibly unknown variance $E\varepsilon_i^2 = \sigma^2 \leq U$.

- Consider isotropic Gaussian design $X_{i,mj} \sim^{i.i.d.} N(0,1)$.
- Or, alternativley, the matrix completion design where the X_i are randomly sampled from $e_k \otimes e_l$, with e_k standard Euclidean basis vectors of $\mathbb{R}^{m_1}, \mathbb{R}^{m_2}$.

- Recall that in such models the minimax recovery rate is

$$r_n^2(R_{m_1 m_2}(k)) = \frac{k(m_1 + m_2)}{n},$$

in Frobenius-norm loss (normalised by $m_1 m_2$ in the matrix completion case).

Confidence sets for low rank matrix recovery II

Theorem (Carpentier, Klopp, Löffler, Nickl)

With unknown variance $\sigma^2 \leq U^2$, the testing rate ρ_n for

$$H_0 : \Theta_0 = R_{m_1 m_2}(k_0) \quad vs. \quad H_1 : \{\theta : \|\theta - \Theta_0\|^2 \geq \rho^2\}$$

satisfies $\rho_n \lesssim (m_1 + m_2)/n$. Since oracle estimators exist, adaptive confidence sets exist that are honest over any pair $\Theta_0 = R_{m_1 m_2}(k_0)$, $\Theta = R_{m_1 m_2}(k)$, $k_0 < k$.

• A simple construction of a confidence region for any adaptive estimator $\hat{\Theta}$ can be given as follows: For matrix completion, if Z_k, Z_k' is any pair of repeated samples of some same matrix entry $\tilde{X}_k$, $k = 1, \dots, N$, form the U-statistic

$$\hat{R} = \frac{1}{N} \sum_{k=1}^{N} (Z_k - \langle \hat{\Theta}, \tilde{X}_k \rangle_F)(Z_k' - \langle \hat{\Theta}, \tilde{X}_k \rangle_F)$$

and take as confidence set

$$C_{n,\alpha} = \left\{ \theta \in \Theta : \frac{\|\theta - \hat{\theta}\|_F^2}{m_1 m_2} \leq \hat{R} + \frac{U^2 + 4\alpha^2}{N} \right\}, \quad 0 < \alpha < 1.$$

Confidence sets for low rank matrix recovery III

$\rightarrow$ The above result is tied to Frobenius norm-loss, and no longer holds true when considering, e.g., the nuclear norm instead (Carpentier & Nickl (2015)).

$\rightarrow$ Moreover, it relies on 'repeated sampling': If one adopts a different formalism for the matrix completion model, namely the 'missing data normal means model'

$$Y_{kl} = (\Theta_{kl} + \varepsilon_{lk})B_{kl},\ 1 \leq k \leq m_1, 1 \leq l \leq m_2,\ B_{kl} \sim^{i.i.d.} \mathrm{Be}(q),\ q = \frac{n}{m_1 m_2} < 1,$$

which is equivalent from a mimimax estimation point of view, then

Theorem (Carpentier, Klopp, Löffler, Nickl)

In matrix completion without repeated sampling, the testing rates satisfy

$$\rho^2 \lesssim \frac{m_1 + m_2}{n} \text{ (known variance)}, \quad \rho_n^2 \gtrsim \frac{\sqrt{k}(m_1 + m_2)}{n} \text{ (unknown variance)}.$$

Thus in the latter case adaptive honest confidence sets do NOT exist over

$$\Theta_0 = R_{m_1 m_2}(k_0), \Theta = R_{m_1 m_2}(k), \quad \text{whenever } k_0 = o(\sqrt{k}).$$

A Bayesian epilogue: Bernstein von Mises theorems in infinite dimensions

- The Bayesian statistician specifies a prior distribution Π on Θ – it encodes prior beliefs about the state of nature. Then one assumes

$$Y|\theta \sim P_\theta$$

and bases its inferences on the 'updated' posterior distribution

$$\theta|Y \sim \Pi(\cdot|Y),$$

derived from Bayes 'theorem'.

A Bayesian epilogue: Bernstein von Mises theorems in infinite dimensions

- The Bayesian statistician specifies a prior distribution Π on Θ – it encodes prior beliefs about the state of nature. Then one assumes

$$Y|\theta \sim P_\theta$$

and bases its inferences on the 'updated' posterior distribution

$$\theta|Y \sim \Pi(\cdot|Y),$$

derived from Bayes 'theorem'.

- Since $\Pi(\cdot|Y)$ is a distribution on Θ it can be used for statistical inference about θ, by computing
  - Bayes decision rules $\bar{\theta}(Y)$ (posterior mean/median)
  - posterior credible regions of $\Pi(\cdot|Y)$-probability 95%, etc.

A basic example: Histogram estimation

- Suppose the $Y_1, \dots Y_n$ are i.i.d. drawn from a density f on $[0,1]$ that is piecewise constant on intervals

$$I_k = \left(\frac{k}{D}, \frac{k+1}{D} \right], \quad k = 0, \dots, D-1.$$

A basic example: Histogram estimation

- Suppose the $Y_1, \dots Y_n$ are i.i.d. drawn from a density f on $[0,1]$ that is piecewise constant on intervals

$$I_k = \left(\frac{k}{D}, \frac{k+1}{D} \right], \quad k = 0, \dots, D-1.$$

- That is, for $\theta = (\theta_k)_{k=0}^{D-1}$ in the unit simplex of $\mathbb{R}^D$, we have

$$f = \sum_{k=0}^{D-1} \theta_k 1_{I_k}.$$

A basic example: Histogram estimation

- Suppose the $Y_1, \dots Y_n$ are i.i.d. drawn from a density f on $[0,1]$ that is piecewise constant on intervals

$$I_k = \left(\frac{k}{D}, \frac{k+1}{D} \right], \quad k = 0, \dots, D-1.$$

- That is, for $\theta = (\theta_k)_{k=0}^{D-1}$ in the unit simplex of $\mathbb{R}^D$, we have

$$f = \sum_{k=0}^{D-1} \theta_k 1_{I_k}.$$

- The MLE equals the proportion of observations falling into I_k:

$$\hat{\theta}_k = \sum_{i=1}^{n} \frac{1_{I_k}(Y_i)}{n}, \quad k = 0, \dots D-1.$$

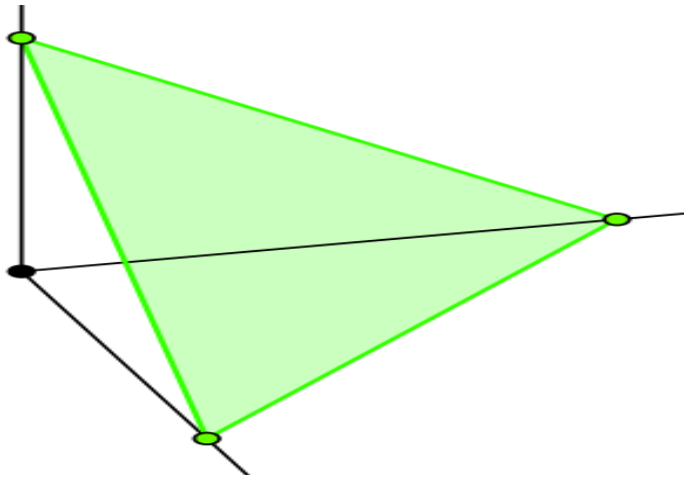
Then $\hat{\theta}_k - \theta_k = \hat{\theta}_k - E_\theta \hat{\theta}_k \sim N(0, 1/(nI_k(\theta)))$ by the CLT.

Bayesian random Dirichlet histograms

Consider a random histogram prior $\Pi = \Pi_L$ defined as

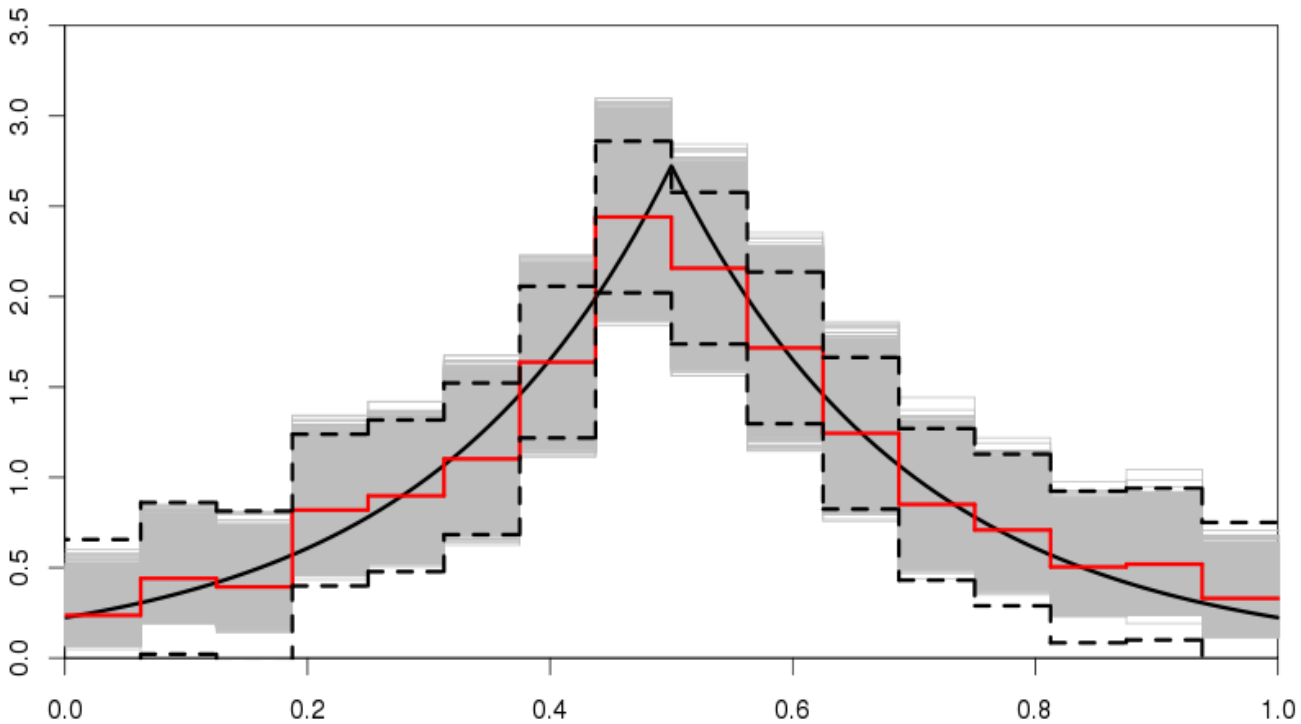
$$f \sim \sum_{k=0}^{D-1} h_k 1_{I_k}$$

with h_k drawn uniformly from the unit simplex of $\mathbb{R}^D$.



Posterior inference

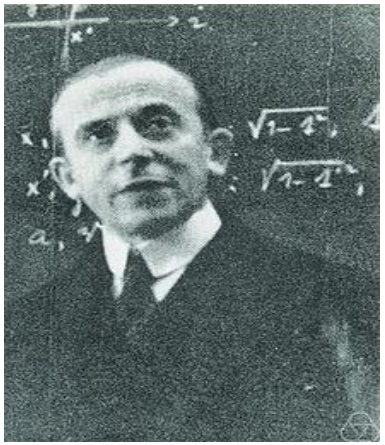
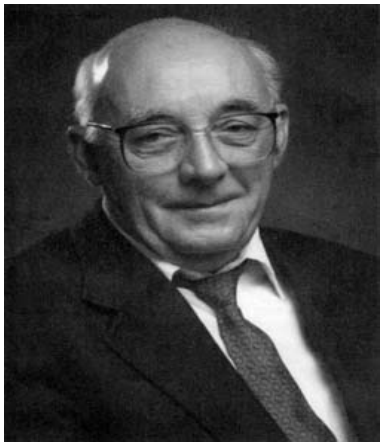
• In this particular case the posterior distribution $\Pi(\cdot|Y_1, \ldots, Y_n)$ can be computed analytically – it is a *Dirichlet distribution*. Posterior mean or median are not equal to the MLE $\hat{\theta}$ (but the posterior mode is).



95%-Euclidean credible ellipsoid (grey) .. Is it a frequentist confidence set?

The Bernstein-von Mises (BvM) theorem



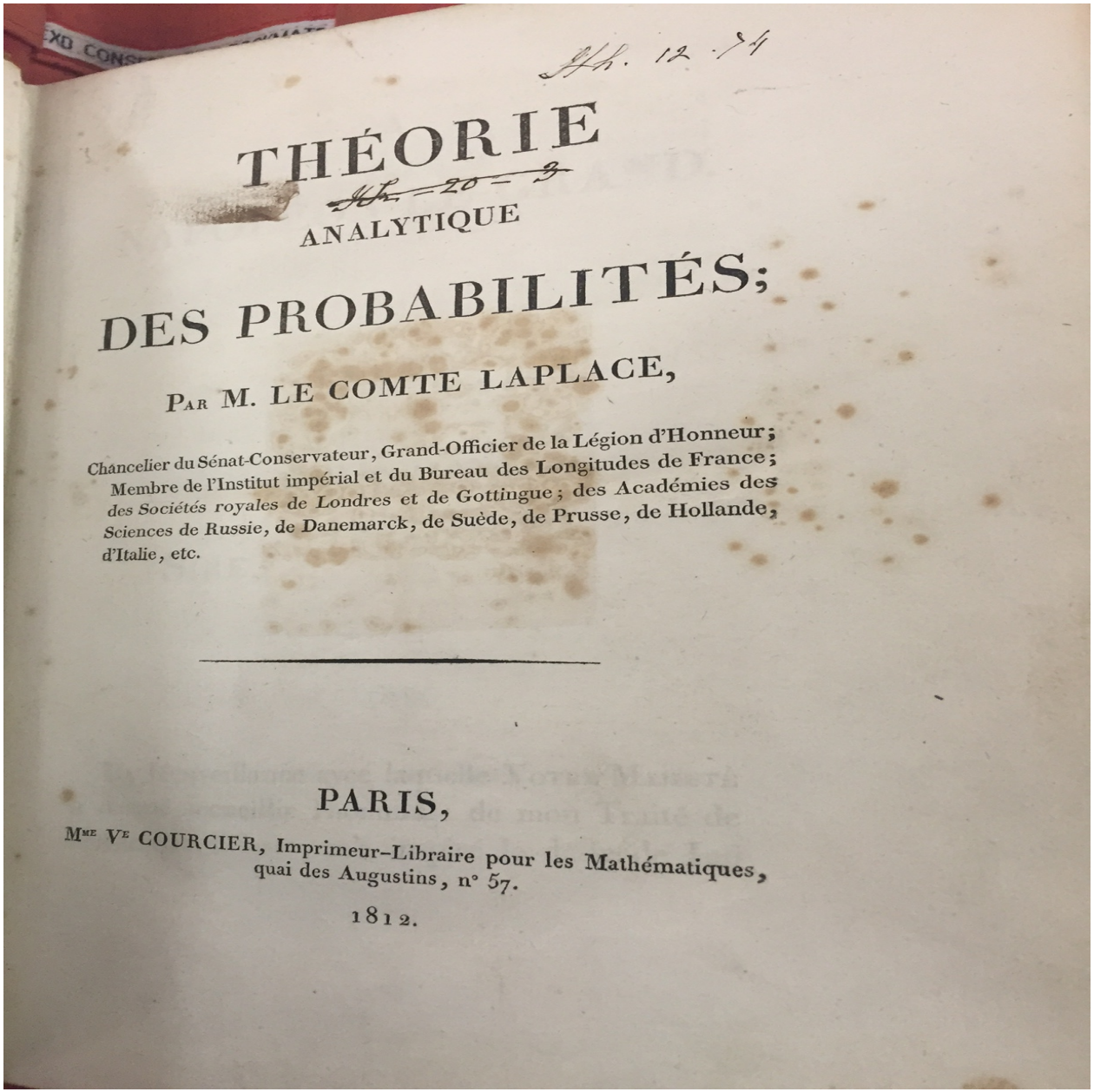



First discovered by P. Laplace (1810), expanded upon by S. Bernstein and R. von Mises in the early 20th century, and proved in its general form by L. Le Cam (1986), the BvM theorem states, for large n:

$$\Pi(\cdot|Y) \approx N(\hat{\theta}_{MLE}, I(\theta_0)^{-1}/n) \text{ for } Y \sim P_{\theta_0}, \theta_0 \in \Theta \subset \mathbb{R}^p$$

whenever the prior has a positive density on Θ. When it exists, the posterior mean $\bar{\theta} = E(\theta|Y)$ can replace $\hat{\theta}_{MLE}$ above.

Laplace 1812



THÉORIE
ANALYTIQUE
DES PROBABILITÉS;

PAR M. LE COMTE LAPLACE,

Chancelier du Sénat-Conservateur, Grand-Officier de la Légion d'Honneur;
Membre de l'Institut impérial et du Bureau des Longitudes de France;
des Sociétés royales de Londres et de Gottingue; des Académies des
Sciences de Russie, de Danemarck, de Suède, de Prusse, de Hollande,
d'Italie, etc.

PARIS,
Mme Ve COURCIER, Imprimeur-Libraire pour les Mathématiques,
quai des Augustins, n° 57.
1812.

A

NAPOLÉON-LE-GRAND.

SIRE,

LA bienveillance avec laquelle VOTRE MAJESTÉ a daigné accueillir l'hommage de mon Traité de Mécanique Céleste, m'a inspiré le desir de Lui

CHAPITRE VI.

De la probabilité des causes et des événemens futurs, tirée des événemens observés.

26. La probabilité de la plupart des événemens simples, est inconnue : en la considérant *à priori*, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais si l'on a observé un résultat composé de plusieurs de ces événemens, la manière dont ils y entrent, rend quelques-unes de ces valeurs plus probables que les autres. Ainsi à mesure que le résultat observé se compose par le développement des événemens simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui se resserrant sans cesse, finiraient par coïncider, si le nombre des événemens simples devenait infini. Pour déterminer les lois suivant lesquelles cette possibilité se découvre, nous la nommerons x. La théorie exposée dans les chapitres précédens, donnera la probabilité du résultat observé, en fonction de x. Soit y cette fonction; si l'on considère les différentes valeurs de x comme autant de causes de ce résultat, la probabilité de x sera, par le troisième principe du n° 1, égale à une fraction dont le numérateur est y, et dont le dénominateur est la somme de toutes les valeurs de y; en multipliant donc le numérateur et le dénominateur de cette fraction par dx, cette probabilité sera

$$\frac{y\,dx}{\int y\,dx},$$

l'intégrale du dénominateur étant prise depuis $x=0$ jusqu'à $x=1$. La probabilité que la valeur de x est comprise dans les limites $x=\theta$ et $x=\theta'$, est par conséquent égale à

$$\frac{\int y\,dx}{\int y\,dx}, \quad (1)$$

v est égal à $\frac{x-a}{\sqrt{\log Y - \log y}}$, et U, $\frac{d.U^2}{dx}$, $\frac{d^2.U^3}{dx^2}$, etc. sont ce que deviennent v, $\frac{d.v^2}{dx}$, $\frac{d^2.v^3}{dx^2}$, etc., lorsqu'on y change après les différentiations, x en a, a étant la valeur de x qui rend y un *maximum*: T' est égal à ce que devient la fonction $\sqrt{\log Y - \log y}$, lorsqu'on change x en $a-\theta$ dans y, et T' est ce que devient la même fonction, lorsqu'on y change x dans $a+\theta'$. L'expression précédente de $\int y\,dx$ donne la valeur de cette intégrale, dans les limites $x=a-\theta$ et $x=a+\theta'$; l'intégrale $\int dt.c^{-t^2}$ étant prise depuis $t=-T$ jusqu'à $t=T'$.

Le plus souvent, aux limites de l'intégrale $\int y\,dx$, étendue depuis $x=0$ jusqu'à $x=1$, y est nul; ou lorsque y n'est pas nul, il devient si petit à ces limites, qu'on peut le supposer nul. Alors, on peut faire à ces limites T et T' infinis, ce qui donne pour l'intégrale $\int y\,dx$, étendue depuis $x=0$ jusqu'à $x=1$,

$$\int y\,dx = Y.\left\{U+\frac{1}{2}.\frac{d^2.U^3}{1.2.dx^2}+\frac{1.3}{2^2}.\frac{d^4.U^5}{1.2.3.4.dx^4}+\text{etc.}\right\}.\sqrt{\pi};$$

ainsi la probabilité que la valeur de x est comprise dans les limites $x=a-\theta$ et $x=a+\theta'$, est égale à

$$\frac{\int dt.c^{-t^2}}{\sqrt{\pi}}+\frac{\left\{\begin{array}{l}\frac{1}{2}.c^{-T^2}.\left\{\frac{d.U^2}{dx}-T.\frac{d^2.U^3}{1.2.dx^2}+(T^2+1).\frac{d^3.U^4}{1.2.3.dx^3}-\text{etc.}\right\}\\ -\frac{1}{2}.c^{-T'^2}.\left\{\frac{d.U^2}{dx}+T'.\frac{d^2.U^3}{1.2.dx^2}+(T'^2+1).\frac{d^3.U^4}{1.2.3.dx^3}+\text{etc.}\right\}\end{array}\right\}}{\left\{U+\frac{1}{2}.\frac{d^2.U^3}{1.2.dx^2}+\frac{1.3}{2^2}.\frac{d^4.U^5}{1.2.3.4.dx^4}+\text{etc.}\right\}.\sqrt{\pi}.} \quad (3)$$

On voit par le n° 23 du premier Livre, que dans le cas où y a pour facteurs, des fonctions de x élevées à de grandes puissances de l'ordre $\frac{1}{\alpha}$, α étant une fraction extrêmement petite, alors U est le plus souvent de l'ordre $\sqrt{\alpha}$, ainsi que ses différences successives; U, $\frac{d.U^2}{dx}$, $\frac{d^2.U^3}{dx^2}$, etc. sont respectivement des ordres $\sqrt{\alpha}$, α, $\alpha^{\frac{3}{2}}$, etc.; d'où il suit que la convergence des séries de la formule (3), exige que T et T' ne soient pas d'un ordre supérieur à $\frac{1}{\sqrt{\alpha}}$.

von Mises 1931

$x_1, x_2, \ldots x_{k+1}$ (also den Grenzwerten der relativen Häufigkeiten) gebildet wird. Man muß nur beachten, daß das α, das in (86) auftritt, nach (70) ebenfalls ein mit den α_ι gebildeter Durchschnittswert ist, während der Mittelwert der Verteilung mit Hilfe der x_ι gerechnet wird. Man nennt passend σ^2 das *„mittlere Abweichungsquadrat der Beobachtungen“* (vgl. dazu § 9, 1). Wohl zu unterscheiden davon ist die Streuung der durch $v_n(x)$ bestimmten Verteilung, die nach der bekannten Eigenschaft der Gaußschen Funktion (§ 2, 4) sich aus (85) zu $\frac{\sigma^2}{n}$ rechnet. Den in (85) enthaltenen *zweiten Fundamentalsatz der Wahrscheinlichkeitsrechnung* sprechen wir wie folgt aus:

Hat die n-malige Beobachtung eines Kollektivs n Resultate ergeben, die den Durchschnitt α und das mittlere Abweichungsquadrat σ^2 aufweisen, so ist die Wahrscheinlichkeitsdichte dafür, daß der Erwartungswert der Verteilung bei x liegt, für hinreichend großes n, gleichgültig wie die Anfangswahrscheinlichkeit beschaffen ist, durch das Gaußsche Gesetz mit dem Mittelwert α und der Streuung $\sigma^2 : n$ gegeben.

Rigorous statements

- Suppose $\Theta \subseteq \mathbb{R}^p$, that the model $\{P_\theta : \theta \in \Theta\}$ is *regular* (LAN) and that the prior has a density that is positive and continuous at $\theta_0 \in int(\Theta)$. Then P_{θ_0}-almost surely,

$$\left\| \mathcal{L}(\sqrt{n}(\theta - \hat{\theta}_{MLE})|Y) - N(0, I(\theta_0)^{-1}) \right\|_{TV} \to 0 \text{ as } n \to \infty,$$

  where $\|\cdot\|_{TV}$ is the total variation distance between two probability distributions. See Le Cam (1986), van der Vaart (1998) for proofs.

Rigorous statements

- Suppose $\Theta \subseteq \mathbb{R}^p$, that the model $\{P_\theta : \theta \in \Theta\}$ is *regular* (LAN) and that the prior has a density that is positive and continuous at $\theta_0 \in int(\Theta)$. Then P_{θ_0}-almost surely,

$$\left\| \mathcal{L}(\sqrt{n}(\theta - \hat{\theta}_{MLE})|Y) - N(0, I(\theta_0)^{-1}) \right\|_{TV} \to 0 \text{ as } n \to \infty,$$

  where $\| \cdot \|_{TV}$ is the total variation distance between two probability distributions. See Le Cam (1986), van der Vaart (1998) for proofs.

- Computing posterior probabilities is then approximately the same as computing them under a $N(\hat{\theta}_{MLE}, I(\theta_0)^{-1}/n)$-distribution, and so:

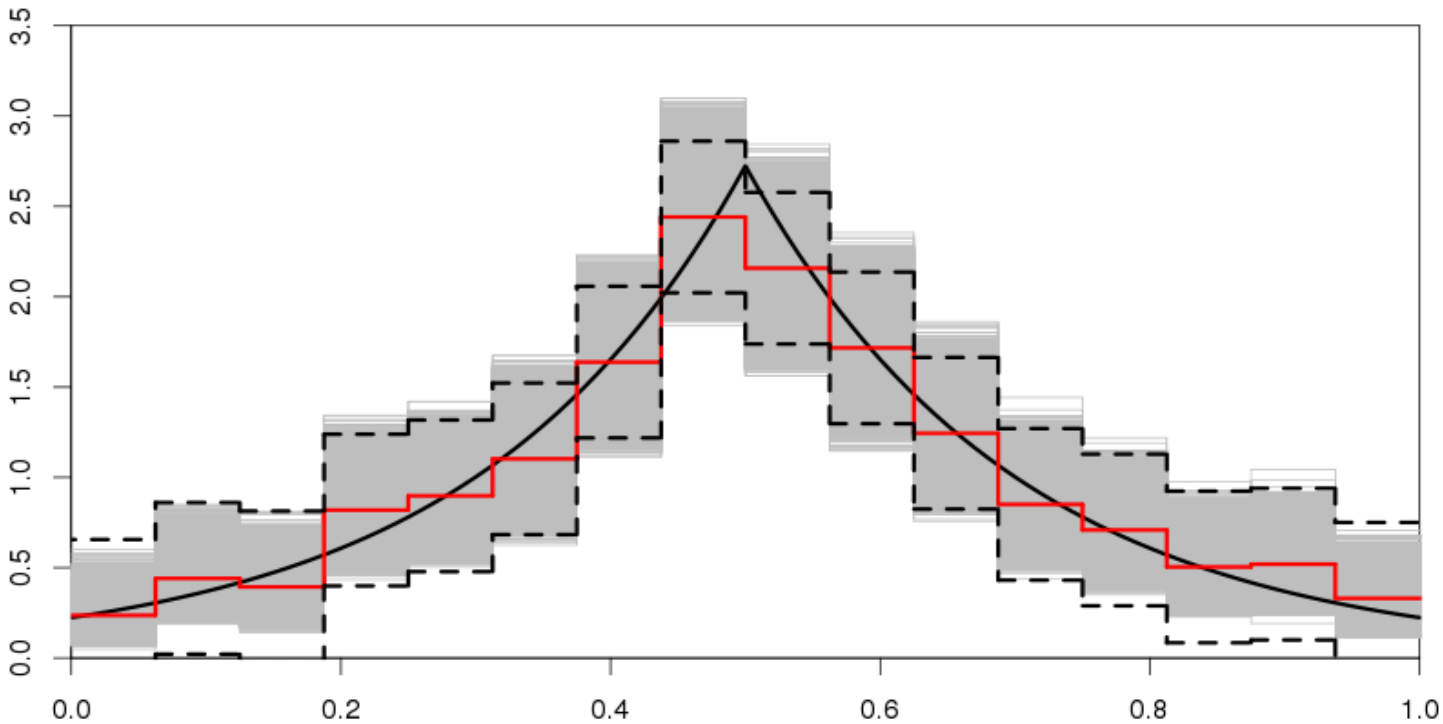
$$C_n \text{ s.t. } \Pi(C_n|Y) = 1 - \alpha \quad \text{(Bayesian credible set)}$$

$$\Rightarrow P_{\theta_0}(\theta_0 \in C_n) \to_{n\to\infty} 1 - \alpha, \quad \text{(frequentist confidence set).}$$

  This is true for C_n of any 'geometry'..!

Application to histograms

• In the 'histogram' example from above, if we assume that the 'true' function f_0 is also piecewise constant on the D bins, then with probability ≈ 0.95 by the D-dimensional Bernstein-von Mises theorem we must have frequentist coverage $\approx P_{f_0}(f_0 \in C_n) \to 0.95$ as $n \to \infty$ of the Bayesian Euclidean credible ball.



Historical summary

→ Laplace and Gauß were both 'Bayesians': they advocated a prior of positive density on the parameter space. They computed 'frequentist' approximations to the posterior mean, or median – mostly for numerical reasons – and found that, surprisingly, this approximation is typically insensitive to the choice of prior.

→ Von Mises realised that the fact that 'any prior washes out in view of the data' was a universal statistical phenomenon.

→ R.A. Fisher, who advocated the use of maximum likelihood methods in the 1920s, started with 'anti-Bayesian' rhetoric, saying that prior choices are arbitrary and resulting inferences unscientific.

→ In finite-dimensional models, Le Cam's work implies that for large sample size, there is NO difference between frequentist and Bayesian inference.

High-dimensional Bernstein – von Mises?

- The BvM theorem is true in finite-dimensional models – in the histogram example when D is fixed and the true parameter θ_0 belongs to the model, that is, if f_0 is itself piecewise constant.

High-dimensional Bernstein – von Mises?

- The BvM theorem is true in finite-dimensional models – in the histogram example when D is fixed and the true parameter θ_0 belongs to the model, that is, if f_0 is itself piecewise constant.
- One may wonder what happens if f_0 is not piecewise constant?

High-dimensional Bernstein – von Mises?

- The BvM theorem is true in finite-dimensional models – in the histogram example when D is fixed and the true parameter θ_0 belongs to the model, that is, if f_0 is itself piecewise constant.
- One may wonder what happens if f_0 is not piecewise constant?
- If only D is large enough, a not too irregular function f_0 would still be well-approximated by a piece-wise constant one. Can we somehow choose the number D 'large', and still make inference on f_0?

High-dimensional Bernstein – von Mises?

- The BvM theorem is true in finite-dimensional models – in the histogram example when D is fixed and the true parameter θ_0 belongs to the model, that is, if f_0 is itself piecewise constant.
- One may wonder what happens if f_0 is not piecewise constant?
- If only D is large enough, a not too irregular function f_0 would still be well-approximated by a piece-wise constant one. Can we somehow choose the number D 'large', and still make inference on f_0?
- This leads to high- and infinite dimensional models, where we think of $D = D_n \to \infty$ with $n \to \infty$.

High-dimensional Bernstein – von Mises?

- The BvM theorem is true in finite-dimensional models – in the histogram example when D is fixed and the true parameter θ_0 belongs to the model, that is, if f_0 is itself piecewise constant.
- One may wonder what happens if f_0 is not piecewise constant?
- If only D is large enough, a not too irregular function f_0 would still be well-approximated by a piece-wise constant one. Can we somehow choose the number D 'large', and still make inference on f_0?
- This leads to high- and infinite dimensional models, where we think of $D = D_n \to \infty$ with $n \to \infty$.
- Does the Bayesian approach still work in such situations?

Nonparametric BvMs I: Diaconis and Freedman (1999)

- For an infinite vector

$$\theta = (\theta_k : k \in \mathbb{N}) \;\; s.t. \;\; \sum_{k \in \mathbb{N}} \theta_k^2 < \infty \iff \theta \in \ell_2,$$

consider observations in the Gaussian sequence space model

$$Y = (Y_k : k \in \mathbb{N}), \;\; Y_k = \theta_k + \frac{1}{\sqrt{n}} g_k, \quad g_k \sim^{i.i.d.} N(0,1).$$

Nonparametric BvMs I: Diaconis and Freedman (1999)

- For an infinite vector

$$\theta = (\theta_k : k \in \mathbb{N}) \quad s.t. \quad \sum_{k \in \mathbb{N}} \theta_k^2 < \infty \iff \theta \in \ell_2,$$

consider observations in the Gaussian sequence space model

$$Y = (Y_k : k \in \mathbb{N}), \quad Y_k = \theta_k + \frac{1}{\sqrt{n}} g_k, \quad g_k \sim^{i.i.d.} N(0,1).$$

- Freedman considered the conjugate situation with Gaussian priors

$$\Pi \sim \bigotimes_{k \in \mathbb{N}} N(0, k^{-(2\gamma+1)}), \quad \gamma > 0.$$

The posterior distribution $\Pi(\cdot|Y)$ is also an infinite Gaussian product measure (so explicit computations are possible in this model).

- Define the posterior mean $\bar{\theta} = E[\theta|Y]$ and the natural ℓ_2-credible ball

$$C_n = \left\{\theta \in \ell_2 : \|\theta - \bar{\theta}\|_{\ell_2}^2 \leq z_{\alpha,n}\right\}$$

where the posterior quantiles $z_{\alpha,n}$ are such that, for some $0 < \alpha < 1$,

$$\Pi(C_n|Y) = 1 - \alpha.$$

- Define the posterior mean $\bar{\theta} = E[\theta|Y]$ and the natural ℓ_2-credible ball

$$C_n = \left\{\theta \in \ell_2 : \|\theta - \bar{\theta}\|_{\ell_2}^2 \leq z_{\alpha,n}\right\}$$

where the posterior quantiles $z_{\alpha,n}$ are such that, for some $0 < \alpha < 1$,

$$\Pi(C_n|Y) = 1 - \alpha.$$

- Let us assume the data are generated from a fixed P_{θ_0}, where θ_0 is γ^*-regular in the sense that is belongs to the ellipsoid

$$\sum_{k\in\mathbb{N}} \theta_{0,k}^2 k^{2\gamma^*} < \infty.$$

If $\gamma < \gamma^*$ we can intuitively think of a 'well specified model' – the true signal is more regular than the typical draw from the prior.

- Diaconis and Freedman proved the following:

- Diaconis and Freedman proved the following:
- There is *no hope* for C_n when $\gamma^* < \gamma$ (a 'misspecified' model.)

- Diaconis and Freedman proved the following:
- There is *no hope* for C_n when $\gamma^* < \gamma$ (a 'misspecified' model.)
- But also when $\gamma^* = \gamma$ or even when $\gamma^* > \gamma$, the pivotal quantity

$$\|\theta - \bar{\theta}\|_{\ell_2}^2, \ \theta \sim \Pi(\cdot|Y),$$

has different Bayesian and frequentist asymptotics:

$$\frac{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|\theta_0)}{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|Y)} \to_{n\to\infty} c < 1.$$

- Diaconis and Freedman proved the following:
- There is *no hope* for C_n when $\gamma^* < \gamma$ (a 'misspecified' model.)
- But also when $\gamma^* = \gamma$ or even when $\gamma^* > \gamma$, the pivotal quantity

$$\|\theta - \bar{\theta}\|_{\ell_2}^2, \ \theta \sim \Pi(\cdot|Y),$$

has different Bayesian and frequentist asymptotics:

$$\frac{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|\theta_0)}{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|Y)} \to_{n\to\infty} c < 1.$$

- Hence a BvM - theorem cannot hold in ℓ_2 as this would force $c = 1$.

- Diaconis and Freedman proved the following:
- There is *no hope* for C_n when $\gamma^* < \gamma$ (a 'misspecified' model.)
- But also when $\gamma^* = \gamma$ or even when $\gamma^* > \gamma$, the pivotal quantity

$$\|\theta - \bar{\theta}\|_{\ell_2}^2, \ \theta \sim \Pi(\cdot|Y),$$

has different Bayesian and frequentist asymptotics:

$$\frac{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|\theta_0)}{Var(\|\bar{\theta} - \theta\|_{\ell_2}^2|Y)} \to_{n\to\infty} c < 1.$$

- Hence a BvM - theorem cannot hold in ℓ_2 as this would force $c = 1$.
- The natural ℓ_2-credible ball centred at the posterior mean is, unlike in the finite-dimensional situation, not a frequentist confidence set:

$$P_{\theta_0}(\theta_0 \in C_n) \nrightarrow 1 - \alpha, \quad \text{possibly} \to 0.$$

Nonparametric BvMs II: The Geometry of Credible Sets

- As before, take a Gaussian prior $\Pi \sim \bigotimes_{k \in \mathbb{N}} N(0, k^{-(2\gamma+1)})$, $\gamma > 0$ on the parameter $\theta \in \ell_2$ of a Gaussian sequence space model.

Nonparametric BvMs II: The Geometry of Credible Sets

- As before, take a Gaussian prior $\Pi \sim \bigotimes_{k\in\mathbb{N}} N(0, k^{-(2\gamma+1)})$, $\gamma > 0$ on the parameter $\theta \in \ell_2$ of a Gaussian sequence space model.
- Assume the model is 'correct': The true coefficients $(\theta_{0,k})$ are s.t.

$$\sup_k \frac{|\theta_{0,k}|}{k^{\gamma+1/2}} < \infty$$

.

Nonparametric BvMs II: The Geometry of Credible Sets

- As before, take a Gaussian prior $\Pi \sim \bigotimes_{k\in\mathbb{N}} N(0, k^{-(2\gamma+1)})$, $\gamma > 0$ on the parameter $\theta \in \ell_2$ of a Gaussian sequence space model.
- Assume the model is 'correct': The true coefficients $(\theta_{0,k})$ are s.t.

$$\sup_k \frac{|\theta_{0,k}|}{k^{\gamma+1/2}} < \infty$$

.

- For any $\delta > 1$, define ellipsoids

$$\mathcal{E}_M = \left\{ (\theta_k) : \sum_{k=1}^{\infty} \frac{\theta_k^2}{k \log^{\delta} k} \leq M^2 \right\}.$$

Nonparametric BvMs II: The Geometry of Credible Sets

- As before, take a Gaussian prior $\Pi \sim \bigotimes_{k\in\mathbb{N}} N(0, k^{-(2\gamma+1)})$, $\gamma > 0$ on the parameter $\theta \in \ell_2$ of a Gaussian sequence space model.
- Assume the model is 'correct': The true coefficients $(\theta_{0,k})$ are s.t.

$$\sup_k \frac{|\theta_{0,k}|}{k^{\gamma+1/2}} < \infty$$

.

- For any $\delta > 1$, define ellipsoids

$$\mathcal{E}_M = \left\{ (\theta_k) : \sum_{k=1}^{\infty} \frac{\theta_k^2}{k \log^{\delta} k} \leq M^2 \right\}.$$

- Define $\mathbb{H}$ to be the normed space that has $\mathcal{E}_1$ as its unit ball. This corresponds to a topology in sequence space that is *weaker* than ℓ_2.

Bernstein von Mises theorem in $\mathbb{H}$

- [Castillo & Nickl 2013 & 2014] Let $\bar{\theta}$ equal the posterior mean $E[\theta|Y]$ or $Y = (Y_k : k \in \mathbb{N})$. As $n \to \infty$ and in P_{θ_0}-probability,

$$\mathcal{L}(\sqrt{n}(\theta - \bar{\theta})|Y) \to \mathcal{N} \text{ weakly in } \mathbb{H},$$

where $\mathcal{N}$ is the Gaussian measure on $\mathbb{H}$ induced by $\otimes_{k\in\mathbb{N}} N(0,1)$.

Bernstein von Mises theorem in $\mathbb{H}$

- [Castillo & Nickl 2013 & 2014] Let $\bar{\theta}$ equal the posterior mean $E[\theta|Y]$ or $Y = (Y_k : k \in \mathbb{N})$. As $n \to \infty$ and in P_{θ_0}-probability,

$$\mathcal{L}(\sqrt{n}(\theta - \bar{\theta})|Y) \to \mathcal{N} \text{ weakly in } \mathbb{H},$$

where $\mathcal{N}$ is the Gaussian measure on $\mathbb{H}$ induced by $\otimes_{k\in\mathbb{N}} N(0,1)$.

- As a consequence, if we define

$$C_n = \left\{\theta : \|\theta - \bar{\theta}\|_{\mathbb{H}} \leq z_{\alpha,n}\right\} \quad \text{with } z_{\alpha,n} \text{ s.t. } \Pi\left(C_n|Y\right) = 1 - \alpha,$$

then one can prove that as $n \to \infty$,

$$P_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha.$$

Bernstein von Mises theorem in $\mathbb{H}$

- [Castillo & Nickl 2013 & 2014] Let $\bar{\theta}$ equal the posterior mean $E[\theta|Y]$ or $Y = (Y_k : k \in \mathbb{N})$. As $n \to \infty$ and in P_{θ_0}-probability,

$$\mathcal{L}(\sqrt{n}(\theta - \bar{\theta})|Y) \to \mathcal{N} \text{ weakly in } \mathbb{H},$$

  where $\mathcal{N}$ is the Gaussian measure on $\mathbb{H}$ induced by $\otimes_{k\in\mathbb{N}} N(0,1)$.

- As a consequence, if we define

$$C_n = \left\{\theta : \|\theta - \bar{\theta}\|_{\mathbb{H}} \leq z_{\alpha,n}\right\} \quad \text{with } z_{\alpha,n} \text{ s.t. } \Pi\left(C_n|Y\right) = 1 - \alpha,$$

  then one can prove that as $n \to \infty$,

$$P_{\theta_0}(\theta_0 \in C_n) \to 1 - \alpha.$$

- An 'adaptive' version of this theorem can also be proved under self-similarity assumptions, see Ray (2017, AoS)

- In nonparametric statistical models, whether a Bayesian credible set is a frequentist confidence set or not may depend, in a possibly nontrivial way, on the geometry of the set.

- In nonparametric statistical models, whether a Bayesian credible set is a frequentist confidence set or not may depend, in a possibly nontrivial way, on the geometry of the set.

- Credible regions arising from ℓ_2-balls are NOT general frequentist confidence sets, whereas $\mathbb{H}$-ball (=ellipsoids) work. The geometry, or topology, is crucial.

- In nonparametric statistical models, whether a Bayesian credible set is a frequentist confidence set or not may depend, in a possibly nontrivial way, on the geometry of the set.

- Credible regions arising from ℓ_2-balls are NOT general frequentist confidence sets, whereas $\mathbb{H}$-ball (=ellipsoids) work. The geometry, or topology, is crucial.

- Analogues of these results also work for the 'histogram example'.

- In nonparametric statistical models, whether a Bayesian credible set is a frequentist confidence set or not may depend, in a possibly nontrivial way, on the geometry of the set.

- Credible regions arising from ℓ_2-balls are NOT general frequentist confidence sets, whereas $\mathbb{H}$-ball (=ellipsoids) work. The geometry, or topology, is crucial.

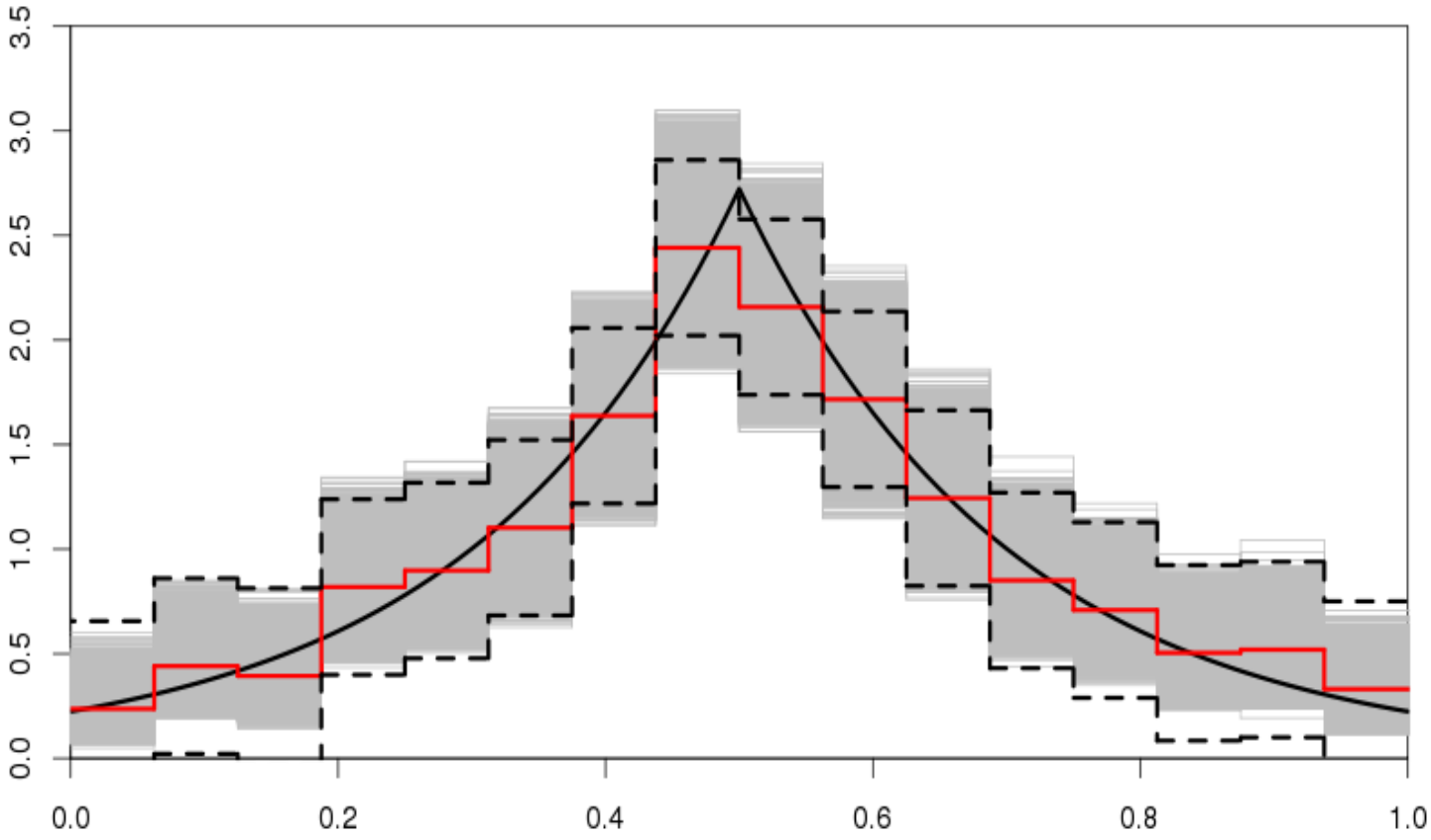
- Analogues of these results also work for the 'histogram example'.

- For instance if we choose D such that $D_n \simeq n^{1/3}$, and f_0 is a truncated Laplace distribution, then $\mathbb{H}$-balls are valid high-dimensional confidence regions.

→ Red line: posterior mean
→ Grey area: $\mathbb{H}$-credible set

→ Black line: true curve f_0
→ Dotted lines: ℓ_2-credible set.

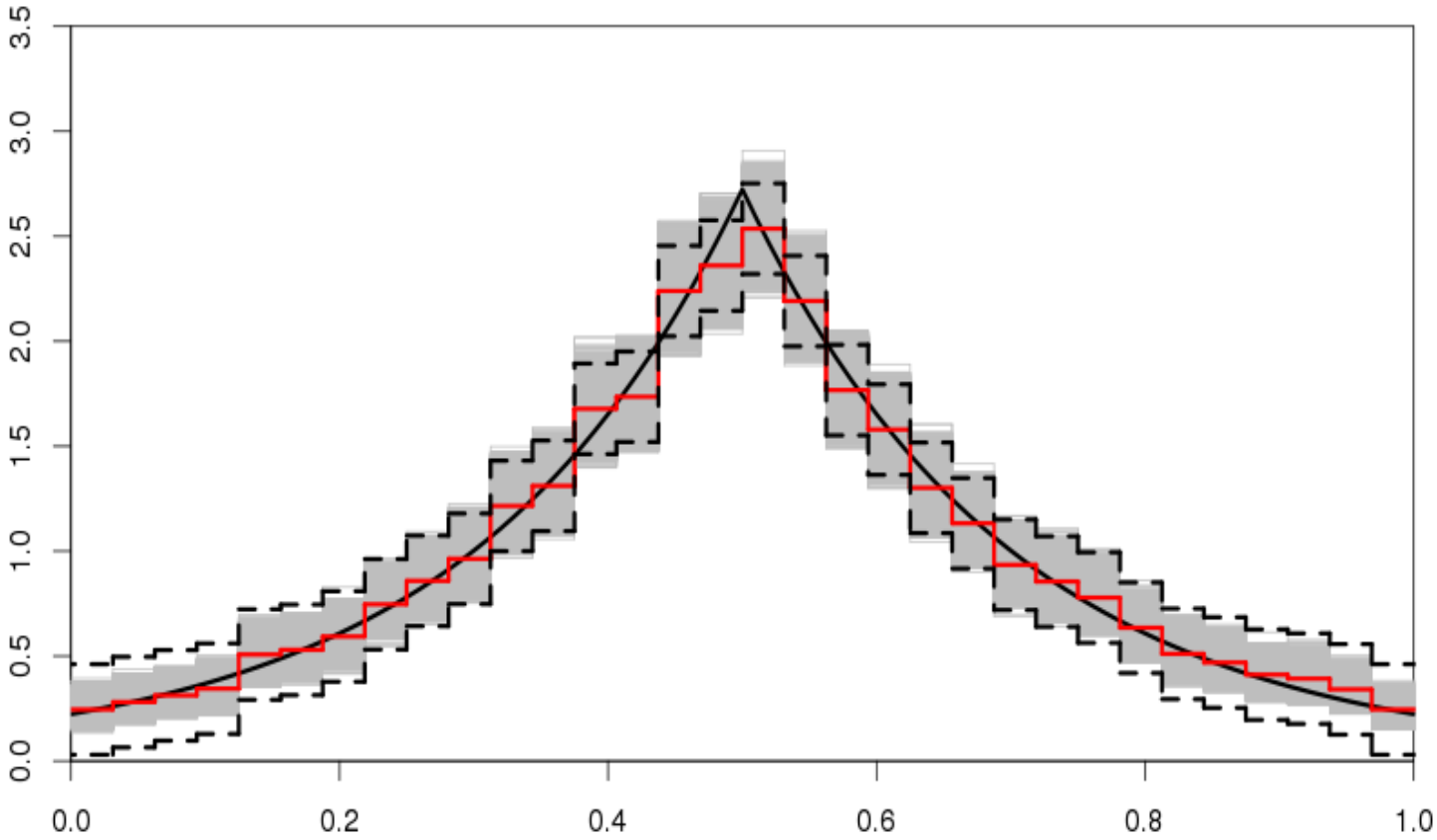
Sample Size $n = 1000$, $\quad f_0(x) \propto e^{-5|x-1/2|} 1_{[0,1]}$



→ Red line: posterior mean → Black line: true curve f_0

→ Grey area: multi-scale credible set → Dotted lines: ℓ_2-credible set.

sample size $n = 10000$, $f_0(x) \propto e^{-5|x-1/2|} 1_{[0,1]}$



Key References:

Bull & Nickl, Adaptive confidence sets in L^2, *Probability Theory Related Fields*, 2013

Cai, Low, Adaptive confidence balls, *Ann. Statist.* 2006

Castillo, Nickl, Nonparametric Bernstein-von Mises theorems in Gaussian white noise, *Ann. Statist.* 2013

Castillo, Nickl, On the Bernstein-von Mises phenomenon for nonparameric Bayes procedures, *Ann. Statist.* 2014

Carpentier, Klopp, Löffler, Nickl, Adaptive confidence sets for matrix completion, *arxiv 2016*

Carpentier, Nickl, On signal detection and low rank inference problems, *EJS*, 2015

Giné & Nickl, Confidence bands in density estimation, *Ann. Statist.*, 2010.

Giné & Nickl, Mathematical foundations of infinite-dimensional statistical models, (Chapter VIII), *Cambridge University Press*, 2016.

Hoffmann & Nickl, On adaptive inference and confidence bands, *Ann. Statist.*, 2011

Low, On nonparametric confidence intervals, *Ann. Statist.*, 1997

Nickl & van de Geer, Confidence sets in sparse regression, *Ann. Statist.*, 2013

Nickl & Szabo, A sharp adaptive confidence ball for self-similar functions, *SPA , 2016*

Robins, van der Vaart, Adaptive nonparametric confidence sets, *Ann. Statist.*, 2006

Szabó, van der Vaart, van Zanten, Frequentist coverage of adaptive nonparametric Bayesian credible sets, *Ann. Statist.*, 2015 (with discussion).