Thanks

Collaborators:
Felix Abramowich
Yoav Benjamini
David Donoho
Noureddine El Karoui
Peter Forrester
Gérard Kerkyacharian
Debashis Paul
Dominique Picard
Bernard Silverman

Presentation help: B. Narasimhan, N. El Karoui, J-M Corcuera

Grant Support: NIH, NSF, ARC (via P. Hall)
Abraham Wald
Three Talks

1. **Function Estimation & Classical Normal Theory**
   - $X_n \sim N_{p(n)}(\theta_n, I)$  \hspace{1cm} $p(n) \nearrow$ with $n$ (MVN)

2. **The Threshold Selection Problem**
   - In (MVN) with, say, $\hat{\theta}_i = X_i I\{|X_i| > \hat{t}\}$
   - How to select $\hat{t} = \hat{t}(X)$ “reliably”?

3. **Large Covariance Matrices**
   - $X_n \sim N_{p(n)}(I \otimes \Sigma_{p(n)})$; especially $X_n = \begin{bmatrix} Y_n \\ Z_n \end{bmatrix}$
   - spectral properties of $n^{-1}X_nX_n^T$
   - PCA, CCA, MANOVA
Focus on Gaussian Models


1943] TESTS OF STATISTICAL HYPOTHESES 433

4. **Reduction of the general problem to the case of a multivariate normal distribution.** In this section we shall prove two lemmas which will enable us to reduce the general problem of large sample inference to the case where the variates under consideration have a joint normal distribution.

**Lemma 1.** *For each positive integer \( n \) there exists a set-function \( W_n^*(W_n) \)

-an inspiration for Le Cam's theory of Local Asymptotic Normality (LAN)
Focus on Gaussian Models, ctd.

• “Growing models”: $p \asymp n$ or $p \gg n$ is now commonplace (classification, genomics..)

• Reality check:

“But, no real data is Gaussian...”

Yes (in many fields), and yet,

consider the power of fable and fairy tale...
1. Function Estimation and Classical Normal Theory

Theme:

Gaussian White Noise model & multiresolution point of view allow classical parametric theory of $N_d(\theta, I)$ to be exploited in nonparametric function estimation.

Reference: book manuscript in (non-)progress:

*Function Estimation and Gaussian Sequence Models*

[www-stat.stanford.edu/~imj](http://www-stat.stanford.edu/~imj)
Agenda

• Classical Parametric Ideas
• Nonparametric Estimation and Growing Gaussian Models
  • I. Kernel Estimation and James-Stein Shrinkage
  • II. Thresholding and Sparsity
  • III. Bernstein-von Mises phenomenon
a) Multinormal shift model

$X_1, \ldots, X_n$ data from $P_\eta(dx) = f_\eta(x)\mu(dx), \quad \eta \in H \subset \mathbb{R}^d$.

Let $I_0 = \text{Fisher information matrix at } \eta_0$.

Local asymptotic Gaussian approximation:

$$\{P^n_{\eta_0 + \theta / \sqrt{n}}, \theta \in \mathbb{R}^d\} \approx \{N_d(\theta, I_0^{-1}), \theta \in \mathbb{R}^d\}$$
b) ANOVA, Projections and Model Selection

\[
Y_{n \times 1} = X_{n \times d} \beta_{d \times 1} + \sigma \epsilon
\]

Submodels: \( X = [X_0 \ X_1] \rightarrow \) projections \( P_{X_0} \).

Canonical form: \( y = \theta + \sigma z \).

Projections:

\[
(P_0 y)_i = \begin{cases} 
  y_i & \text{if } i \in I_0 \\
  0 & \text{o/w}
\end{cases}
\]

MSE:

\[
E \| P_0 y - \theta \|^2 = \sum_{i \in I_0} \sigma^2 + \sum_{i \notin I_0} \theta_i^2
\]
Classical Ideas, ctd

c) Minimax estimation of $\theta$

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}^d} E_{\theta} \| \hat{\theta}(y) - \theta \|^2 = d\sigma^2$$

attained by MLE $\hat{\theta}_{MLE}(y) = y$.

d) James-Stein estimate

$$\hat{\theta}^{JS}(y) = \left(1 - \frac{d - 2}{\|y\|^2}\right)y$$

dominates MLE if $d \geq 3$. 
Classical Ideas, ctd

e) Conjugate priors
Prior: \( \theta \sim N(0, \tau^2 I) \), Likelihood: \( y|\theta \sim N(\theta, \sigma^2 I) \)
Posterior: \( \theta|y \sim N(\hat{\theta}_y, \sigma_y^2) \)

\[
\hat{\theta}_y = \frac{\tau^2}{\sigma^2 + \tau^2 y} \quad \sigma_y^2 = \frac{\sigma^2 \tau^2}{\sigma^2 + \tau^2}
\]

f) Unbiased risk estimate
\( Y \sim N_d(\theta, I) \). For \( g \) (weakly) differentiable

\[
E\|Y + g(Y) - \theta\|^2 = E[d + 2\nabla^T g(Y) + \|g(Y)\|^2] = E[U(Y)]
\]

If \( g = g(Y; t) \) then \( U = U(Y; t) \).
Agenda

- Classical Parametric Ideas
- **Nonparametric Estimation and Growing Gaussian Models**
  - I. Kernel Estimation and James-Stein Shrinkage
  - II. Thresholding and Sparsity
  - III. Bernstein-von Mises phenomenon
Nonparametric function estimation

- Hodges and Fix (1951) nonparametric classification
- spectrum estimation in time series,
- kernel methods for density est’n, regression
- roughness penalty methods (splines)
- techniques largely differ from parametric normal theory

- Ibragimov & Hasminskii (and school): importance of Gaussian white noise (GWN) model:

\[ Y_t = \int_0^t f(s) ds + \epsilon W_t, \quad 0 \leq t \leq 1 \]
Motivation for GWN model

GWN model emerges as large-sample limit of *equispaced regression*, density estimation, spectrum estimation, ...

\[
y_j = f(j/n) + \sigma w_j \\
j = 1, \ldots, n
\]

\[
\epsilon = \sigma / \sqrt{n}
\]
Motivation for GWN model

GWN model emerges as large-sample limit of *equispaced regression*, density estimation, spectrum estimation,...

\[ y_j = f\left(\frac{j}{n}\right) + \sigma w_j \]
\[ j = 1, \ldots, n \]
\[ \epsilon = \frac{\sigma}{\sqrt{n}} \]

\[
\frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} y_j = \frac{1}{n} \sum_{j=1}^{\lfloor nt \rfloor} f\left(\frac{j}{n}\right) + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor nt \rfloor} w_j
\]
Motivation for GWN model

GWN model emerges as large-sample limit of *equispaced regression*, density estimation, spectrum estimation,…

\[
y_j = f\left(\frac{j}{n}\right) + \sigma w_j
\]

\[
j = 1, \ldots, n
\]

\[
\epsilon = \sigma / \sqrt{n}
\]

\[
\frac{1}{n} \sum_{j=1}^{[nt]} y_j = \frac{1}{n} \sum_{j=1}^{[nt]} f\left(\frac{j}{n}\right) + \frac{\sigma}{\sqrt{n}} \frac{1}{\sqrt{n}} \sum_{j=1}^{[nt]} w_j
\]

\[
Y_t = \int_0^t f(s)ds + \frac{\sigma}{\sqrt{n}} \cdot W_t
\]
Series form of WN Model

\[ Y_t = \int_0^t f(s) ds + \epsilon W_t \]

For any orthonormal basis \( \{ \psi_\lambda, \lambda \in \Lambda \} \) for \( L_2[0, 1] \),

\[ \int \psi_\lambda dY_t = \int \psi_\lambda f dt + \epsilon \int \psi_\lambda dW_t \]

\[ \Rightarrow y_\lambda = \theta_\lambda + \epsilon z_\lambda, \quad (z_\lambda) \text{i.i.d. } \sim N(0, 1), \quad \lambda \in \Lambda \]

Parseval relation implies

\[ \int (\hat{f} - f)^2 = \sum (\hat{\theta}_\lambda - \theta_\lambda)^2 = \| \hat{\theta} - \theta \|^2, \quad \text{etc.} \]

→ analysis of infinite sequences in \( \ell_2(\mathbb{N}) \)
Wavelet orthonormal bases \( \{ \psi_{jk} \} \) have \textbf{double} index

- level ("octave") \( j = 1, 2, \ldots \)
- location \( k = 1, \ldots, 2^j \)

Collect coefficients in a single level:

\[
y_j = (y_{jk}) \quad k = 1, \ldots, 2^j
\]

\[
\theta_j = (\theta_{jk}) \quad \text{etc.}
\]

\textbf{Finite} (but growing with \( j \)) multivariate normal model

\[
y_j \sim N_{2^j}(\theta_j, \epsilon^2 I), \quad j = 1, 2, \ldots
\]

\( \Rightarrow \) apply classical normal theory to the vectors \( y_j \).
Some S8 Symmlets at various scales and locations

(8,202)
(8,166)
(8,102)
(8,81)
(8,31)
(7,101)
(7,77)
(7,51)
(6,42)
(6,34)
(6,26)
(6,12)
(4,11)
(4,8)
(4,3)
Example

- Nationwide electricity consumption in France
- Here: 3 days (Thur, Fri, Sat), summer 1990, $\Delta t = 1$ min.
- N.B. Malfunction of meters: d. 2, 16.30 – d. 3, 09.00
Often o.k. to treat a single level as $\sim N_{2^j}(\theta_j, \epsilon^2 I)$
Or, apply $N_d(\theta, I)$ model to a block within a level.
The Minimax principle

Given loss function \( L(a, \theta) \), and risk

\[
R(\hat{\theta}, \theta) = E_\theta L(\hat{\theta}, \theta)
\]

choose estimator \( \hat{\theta} \) so as to minimize the \textit{maximum risk}

\[
\sup_{\theta \in \Theta} E_\theta L(\hat{\theta}, \theta).
\]

• introduced into statistics by Wald
• L.J. Savage (1951):
  “the only rule of comparable generality proposed since [that of] Bayes’ was published in 1763”
• standard complaint: \( \hat{\theta} \) depends on \( \Theta \):
  • optimizing on the worst case may be irrelevant
Simultaneous near-minimaxity

Shift in perspective on minimaxity

- fix $\hat{f}$ in advance; an estimator to be evaluated
- for many spaces $\mathcal{F}$, compare

$$\sup_{\mathcal{F}} R(\hat{f}, f) \quad \text{to} \quad R(\mathcal{F}) = \inf_{\hat{f}} \sup_{\mathcal{F}} R(\hat{f}, f)$$

- shift from

  “Exact answer to ‘wrong’ problem”

  to

  “Approx answer to many (related) problems”

$\Rightarrow$ an imperfect, yet serviceable tool
Agenda

• Classical Parametric Ideas
• Nonparametric Estimation and Growing Gaussian Models
  • I. Kernel Estimation and James-Stein Shrinkage
  • II. Thresholding and Sparsity
  • III. Bernstein-von Mises phenomenon
I. Kernel estimation & James-Stein Shrinkage

Priestley-Chao estimator:

\[ \hat{f}_h(t) = \frac{1}{Nh} \sum_{i=1}^{N} K \left( \frac{t - t_i}{h} \right) Y_i \]

\text{e.g. } K(x) = (1 - x^2)^4

- Automatic choice of \( h \)?
- Huge literature, (e.g. Wand & Jones, 1995)
- James-Stein provides simple, powerful approach
Fourier Form of Kernel Smoothing

Convolution $\rightarrow$ multiplication

Shrinkage factors
- $s_k(h) \in [0, 1]$
- decrease with frequency

Flatten shrinkage in blocks:

$$\hat{\theta}_k = (1 - w_j)y_k \quad k \in B_j$$

- How to choose $w_j$?
James-Stein Shrinkage

For $y \sim N_d(\theta, I)$, and $d \geq 3$

$$\hat{\theta}^{JS+} = \left(1 - \frac{d - 2 + \eta}{|y|^2}\right) y$$

Unbiased estimate of risk:

$$E\|\hat{\theta}^{JS} - \theta\|^2 = E_\theta \left\{d - \frac{(d - 2)^2 + \eta^2}{\|y\|^2}\right\} \leq d$$
Smoothing via Blockwise James-Stein

Block Fourier or wavelet (here) coefficients:

\[ y_j = (y_{jk}), \quad k = 1, \ldots, d_j = 2^j; \quad \theta_j = (\theta_{jk}) \] etc.

Apply J-S shrinkage to each block: \[ \text{refs} \]

\[ \hat{\theta}_{JS}^+ = \left(1 - \frac{\beta_j \epsilon^2}{|y_j|^2}\right) y_j \quad 2 \leq j \leq \log_2 N \]

\[ \beta_j = \begin{cases} 
  d_j - 2 & (\text{ExactJS}) \text{ or} \\
  d_j + \sqrt{2d_j} \sqrt{2 \log d_j} & (\text{ConsJS})
\end{cases} \]

For ConsJS, if \( \theta_j = 0 \), then \( P\{|y_j|^2 < \beta_j\} \nearrow 1 \), so \( \hat{\theta}_{JS}^+ = 0 \).
Block James-Stein imitates Kernel
Block James-Stein imitates Kernel

Red: James–Stein equivalent kernels; Blue: Actual Kernel
Properties of Block James-Stein

• Block JS imitates kernel smoothing, **but**
• near automatic choice of ’bandwidth’
  – canonical choices of $\beta_j$
• Easy theoretical analysis from unbiased risk bounds

$$E \| \hat{\theta}_j^{JS} - \theta_j \|^2 \leq 2\varepsilon^2 + \| \theta_j \|^2 \land 2^j \varepsilon^2$$
Properties of Block James-Stein

- Block JS imitates kernel smoothing, but
- near automatic choice of 'bandwidth'
  – canonical choices of $\beta_j$
- Easy theoretical analysis from unbiased risk bounds
  \[ E\|\hat{\theta}_j^{JS} - \theta_j\|^2 \leq 2\epsilon^2 + \|\theta_j\|^2 \land 2^j \epsilon^2 \]
- e.g. below: MSE for Hölder smooth functions:
  - $0 < \delta < 1 \quad |f(x) - f(y)| \leq B|x - y|^\delta$  all $x, y$ (*)
  - $\alpha = r + \delta, r \in \mathbb{N}, \quad D^r f$ satisfies (*).
- In terms of wavelet coefficients:
  \[ f \in \mathcal{H}_\alpha(C') \iff |\theta_{jk}| \leq C 2^{-(\alpha+1/2)j} \quad \text{for all } j, k \]
A single level determines MSE

From unbiased risk bounds and Hölder smoothness:

\[ E\|\hat{\theta}_j^{JS} - \theta_j\|^2 \leq [2\epsilon^2 + \|\theta_j\|^2 \wedge 2^j \epsilon^2] \]

\[ f \in \mathcal{H}_\alpha(C) \Rightarrow \|\theta_j\|^2 \leq C^{2-2\alpha j} \]
A single level determines MSE

From unbiased risk bounds and Hölder smoothness:

\[
\sum_j E\|\hat{\theta}_{JS}^j - \theta_j\|^2 \leq \sum_{j \leq J} [2\epsilon^2 + \|\theta_j\|^2 \land 2^j \epsilon^2] + \sum_{j > J} \|\theta_j\|^2
\]

\[f \in \mathcal{H}_\alpha(C) \Rightarrow \|\theta_j\|^2 \leq C^2 2^{-2\alpha j}\]
Growing Gaussians

- geometric decay of MSE away from critical $j_*$
- **Growing Gaussian** aspect: As noise $\epsilon = \sigma / \sqrt{n}$ decreases, worst level $j_* = j_*(\epsilon, C)$ increases:

$$d_{j_*} = 2^{j_*} = c(C/\epsilon)^{2/(2\alpha+1)}$$
MultiJS is rate-adaptive

The “optimal rate” for $\mathcal{H}_\alpha(C)$ is $\epsilon^{2r}$, with $r = r(\alpha) = \frac{2\alpha}{2\alpha+1}$.

- JS risk bounds imply simultaneous near-minimality:

$$\sup_{\mathcal{H}_\alpha(C)} R(\hat{f}^{JS}, f) \leq cC^{2(1-r)} \epsilon^{2r} (1 + o(1)) \asymp R(\mathcal{H}_\alpha(C)).$$

- For a **single, prespecified** estimator $\hat{f}^{JS}$, valid for
  - all smoothness $\alpha \in (0, \infty)$,
  - bounds $C \in (0, \infty)$.

- No “speed limit” to rate of convergence
  - “infinite order kernel” (for certain wavelets)

- Conclusion follows easily from single level James-Stein analysis in multinormal mean model
Agenda

- Classical Parametric Ideas
- Nonparametric Estimation and Growing Gaussian Models
  - I. Kernel Estimation and James-Stein Shrinkage
  - II. Thresholding and Sparsity
  - III. Bernstein-von Mises phenomenon
James-Stein Fails on Sparse Signals

\[
\frac{d\|\theta\|^2}{d + \|\theta\|^2} \leq R(\hat{\theta}^{JS}, \theta) \leq 2 + \frac{d\|\theta\|^2}{d + \|\theta\|^2}
\]

\(r\text{-spike } \theta_r: \text{ } r \text{ co-ords at } \sqrt{d/r} \Rightarrow \|\theta\|^2 = d.

So \(R(\hat{\theta}^{JS}, \theta_r) \geq d/2\)
James-Stein Fails on Sparse Signals

\[
\frac{d \| \theta \|^2}{d + \| \theta \|^2} \leq R(\hat{\theta}^{JS}, \theta) \leq 2 + \frac{d \| \theta \|^2}{d + \| \theta \|^2}
\]

\(r\)-spike \(\theta_r\): \(r\) co-ords at \(\sqrt{d/r}\) \(\Rightarrow\) \(\| \theta \|^2 = d\).

So \(R(\hat{\theta}^{JS}, \theta_r) \geq d/2\)

**Hard thresholding** at \(t = \sqrt{2 \log d}\):

\[
R(\hat{\theta}^{HT}, \theta_r) \approx d \cdot 2t \phi(t) + r \cdot [1 + \eta(d/r)] \\
\leq r + 3\sqrt{2 \log d}
\]
A more systematic story for thresholding

Based on $\ell_p$ norms and balls: (mostly with $p < 2$)

$$\|\theta\|^p_p = \sum_{i=1}^{d} |\theta_i|^p$$

1. $\ell_p$ norms capture sparsity
2. thresholding arises from least squares estimation with $\ell_p$ constraints
3. describe best possible estimation over $\ell_p$ balls
4. show that thresholding (nearly) attains the best possible
$\ell_p$ norms, $p < 2$, capture sparsity

$\ell_1$ ball smaller than $\ell_2$, $\Rightarrow$ expect better estimation
\( \ell_p \) norms, \( p < 2 \), capture sparsity

\[ \ell_1 \text{ ball smaller than } \ell_2, \quad \Rightarrow \text{expect better estimation} \]

Sparsity of representation depends on basis: rotation sends

\[ (0, C, 0, \ldots) \rightarrow \left( \frac{C}{\sqrt{d}}, \ldots, \frac{C}{\sqrt{d}} \right) \]
Thresholding from constrained least squares

\[
\min \sum (y_i - \theta_i)^2 \quad \text{s.t.} \quad \sum |\theta_i|^p \leq C^p
\]

i.e.
\[
\min \sum (y_i - \theta_i)^2 + \lambda |\theta_i|^p
\]

leads to

- \( p = 2 \) **Linear Shrinkage** \( \hat{\theta}_i = (1 + \lambda)^{-1} y_i \)

- \( p = 1 \) **Soft thresholding** \( \hat{\theta}_i = \begin{cases} y_i - \lambda' & y_i > \lambda' \\ 0 & |y_i| \leq \lambda' \\ y_i + \lambda' & y_i < -\lambda' \end{cases} \) \( [\lambda' = \lambda/2] \)

- \( p = 0 \) **Hard thresholding** \( \hat{\theta}_i = y_i I\{|y_i| > \lambda\} \).
  \[ \text{[penalty } \lambda \sum I\{|\theta_i| \neq 0\} \]
Bounds on best estimation on $\ell_p$ balls

Minimax risk: $R_{d,p}(C) = \inf_{\hat{\theta}} \sup_{\|\theta\|_p \leq C} E_{\theta} \sum_1^d (\hat{\theta}_i - \theta_i)^2$.

Non-asymptotic bounds: [Birgé-Massart]

$$c_1 r_{d,p}(C') \leq R_{d,p}(C) \leq c_2 [\log d + r_{d,p}(C')]$$

- $\ell_1$ ball smaller $\rightarrow$ (much) smaller minimax risk
Bounds on best estimation on $\ell_p$ balls

Minimax risk:  
$$R_{d,p}(C) = \inf_\hat{\theta} \sup_{\|\theta\|_p \leq C} E_\theta \sum_1^d (\hat{\theta}_i - \theta_i)^2.$$

Non-asymptotic bounds:  
[ Birgé-Massart]  
$$c_1 r_{d,p}(C') \leq R_{d,p}(C) \leq c_2 [\log d + r_{d,p}(C')]$$

- $\ell_1$ ball smaller $\rightarrow$ (much) smaller minimax risk
Thresholding nearly attains the bound

For simplicity: threshold $t = \sqrt{2 \log d}$

Oracle inequality for soft thresholding: \textit{[non-asymptotic!]}\]

$$E\|\hat{\theta}^{ST} - \theta\|^2 \leq (2 \log d + 1)[1 + \sum(\theta_i^2 \wedge 1)]$$

Apply to $\ell_1$ ball $\{\theta : \|\theta\|_1 \leq C\}$.

$$\sup_{\|\theta\|_1 \leq C} E\|\hat{\theta}^{ST} - \theta\|^2 \leq (2 \log d + 1)(C + 1).$$

Better bounds possible for data-dependent thresholds $t = t(Y)$

(Lec. 2)
Summary for $N_d(\theta, I)$

For $X \sim N_d(\theta, I)$, comparing

- James Stein shrinkage ($\beta = d - 2$), and
- soft thresholding at $t = \sqrt{2 \log d}$:

James-Stein shrinkage: orthogonally invariant:

$$\frac{1}{2}(\|\theta\|^2 \wedge d) \leq E\|\hat{\theta}^{JS} - \theta\|^2 \leq 2 + (\|\theta\|^2 \wedge d)$$

Thresholding: co-ordinate wise, and co-ordinate dependent:

$$\frac{1}{2} \sum (\theta_i^2 \wedge 1) \leq E\|\hat{\theta}^{ST} - \theta\|^2 \leq (2 \log d + 1)(1 + \sum (\theta_i^2 \wedge 1))$$
Implications for Function Estimation

Key example: functions of bounded total variation:

$$TV(C) = \{ f : \| f \|_{TV} \leq C \}$$

$$\| f \|_{TV} = \sup_{t_1 < \ldots < t_N} \sum_{i=1}^{N-1} |f(t_{i+1}) - f(t_i)| + \| f \|_1$$

Well captured by weighted combinations of $\ell_1$ norms on wavelet coefficients:

$$c_1 \sup_j 2^{j/2} \| \theta_j \|_1 \leq \| f \|_{TV} \leq c_2 \sum_j 2^{j/2} \| \theta_j \|_1$$

Best possible (minimax) MSE

$$R(TV(C), \epsilon) = \inf_f \sup_{f \in TV(C)} E \| \hat{f} - f \|_2$$

- p. 43
Reduction to single levels

\[
\sup_{\theta} E\|\hat{\theta} - \theta\|^2 = \sum_{j} \sup_{\theta_j} E\|\hat{\theta}_j - \theta_j\|^2
\]

- Apply *thresholding* bounds for each \( j \): to \( N_{2j}(\theta_j, \epsilon^2 I) \).
- \( \exists \) a worst case level \( j^* = j^*(\epsilon, C') \),
- Geometric decay of \( \sup_j E\|\hat{\theta}_j - \theta_j\|^2 \)
  as \( |j - j^*| \nearrow \).

**Growing Gaussian** aspect: As noise \( \epsilon = \sigma/\sqrt{n} \) decreases,
worst level \( j^* = j^*(\epsilon, C') \) increases:

\[
d_{j^*} = 2^{j^*} = c(C/\epsilon)^{2/3}
\]
Final Result

\[
\sup_{TV(C)} R(\hat{f}_{thr}, f) \asymp C^{2(1-r)} \epsilon^{2r} \quad r = 2/3
\]

\[
\sup_{TV(C)} R(\hat{f}_{JS}, f) \asymp C^{2(1-r_L)} \epsilon^{2r_L} \quad r_L = 1/2
\]
Agenda

• Classical Parametric Ideas
• Nonparametric Estimation and Growing Gaussian Models
  • I. Kernel Estimation and James-Stein Shrinkage
  • II. Thresholding and Sparsity
• III. Bernstein-von Mises phenomenon
Bernstein-von Mises Phenomenon

Asymptotic match of freq. & Bayesian confidence intervals:

Classical version

\[ Y_1, \ldots, Y_n \stackrel{i.i.d.}{\sim} p_\theta(y) d\mu \quad \theta \in \Theta \subset \mathbb{R}^d \quad \text{d fixed} \]

Under mild conditions, as \( n \to \infty \),

\[ \left\| P_{\theta|Y} - N(\hat{\theta}_{MLE}, n^{-1} I^{-1}_{\theta_0}) \right\| \xrightarrow{P_{n,\theta_0}} 0, \]

where

- \( I_\theta = E_\theta \left[ \frac{\partial}{\partial \theta} \log p_\theta \right] \left[ \frac{\partial}{\partial \theta} \log p_\theta \right]^T \) is Fisher information,
- \( \|P - Q\| = \max_A |P(A) - Q(A)| \) is total variation distance
Non-parametric Regression

\[ Y_i = f(i/n) + \sigma_0 w_i, \quad i = 1, \ldots, n \]

For typical smoothness priors (e.g. \(d^{th}\) integrated Wiener process prior)

- Bernstein - von Mises fails [Dennis Cox, D. A. Freedman]
- frequentist & posterior laws of \(\|\hat{f} - f\|_2^2\) mismatch in center and scale

Here,

- revisit via elementary Growing Gaussian Model approach
- Apply to single levels of wavelet decomposition
Growing Gaussian Model

Data: \( Y_1, \ldots, Y_n | \theta \) i.i.d. \( N_p(\theta, \sigma_0^2 I) \)  

Prior: \( \theta \sim N_p(0, \tau_n^2) \).

- growing: \( p = p(n) \uparrow \) with \( n \)
- \( \sigma_n^2 = \text{Var}(\bar{Y}_n|\theta) = \sigma_0^2/n; \quad \tau_n^2 \) may depend on \( n \).

Goal: compare \( \mathcal{L}(\hat{\theta}_{MLE}|\theta), \mathcal{L}(\hat{\theta}_{Bayes}|\theta) \) with \( \mathcal{L}(\theta|Y) \).

Posterior: All Gaussian: centering and scaling are key:

\[
\mathcal{L}(\theta|\bar{Y} = \bar{y}) \sim N_p\left( \hat{\theta}_B = w_n \bar{y}, \ w_n \sigma_n^2 I \right)
\]

\[
w_n = \frac{\tau_n^2}{\sigma_n^2 + \tau_n^2}
\]
Growing Gaussians and BvM

Correspondences:

\[ P_{\theta|Y} \leftrightarrow N_p(w_n \bar{y}, \ w_n \sigma_n^2 I) \]
\[ N\left(\hat{\theta}_{MLE}, n^{-1}I^{-1}_{\theta_0}\right) \leftrightarrow N_p(\bar{y}, \sigma_n^2 I) \]

For BvM to hold, now need \( w_n \nrightarrow 1 \) sufficiently fast:

**Proposition:** In the growing Gaussian model

\[
\| P_{\theta|Y} - N\left(\hat{\theta}_{MLE}, n^{-1}I^{-1}_{\theta_0}\right) \| \xrightarrow{P_{n,\theta_0}} 0,
\]

if and only if

\[
\sqrt{p} \frac{\sigma_n^2}{\tau_n^2} = \sqrt{p} \frac{\sigma_0^2}{n \tau_n^2} \to 0 \quad \text{i.e.} \quad w_n = 1 - o\left(\frac{1}{\sqrt{p_n}}\right).
\]
Example: Pinsker priors

Back to regression: \[ dY_t = f(t)dt + \sigma_n dW_t \]

Minimax MSE linear estimation of \( f \):
\[
\left\{ f : \int_0^1 (D^\alpha f)^2 \leq C^2 \right\}
\]

Least favourable prior on wavelet coeffs (for sample size \( n \)):

\[
\theta_{jk} \overset{\text{indep}}{\sim} N(0, \tau_j^2)
\]

\[
\tau_j^2 = \sigma_n^2 (\lambda_n 2^{-j\alpha} - 1)_+
\]

\[
\lambda_n = c (C/\sigma_n)^{2\alpha/(2\alpha+1)}
\]

\Rightarrow \text{critical level } j_* = j_*(n, \alpha) \text{ grows with } n

(Growing Gaussian model again)
Validity of B-vM depends on level

Bayes estimator for the Pinsker prior attains exact minimax MSE (asymptotically)

**But** Bernstein von Mises fails at the critical level $j^*(n)$:

At $j^*(n)$

\[
\frac{\tau_{j^*}^2}{\sigma_n^2} \leq 2^\alpha - 1
\]

[fine scale features]

\[
1 - w_n = \frac{1}{1 + \frac{\tau_{j^*}^2}{\sigma_n^2}} \geq 2^{-\alpha} \quad \text{BvM fails}
\]

At fixed $j_0$:

\[
\begin{align*}
p &= 2^{j_0} \quad \text{FIXED} \quad \text{(or slowly $\nearrow$)} \\
\frac{\tau_{j_0}^2}{\sigma_n^2} &= \lambda_n 2^{-j_0 \alpha} & \to \infty \\
1 - w_n &\to 0 \\
\end{align*}
\]

BvM holds

$\Rightarrow$ Difficulty lies with high dimensional features.
Three Talks

1. **Function Estimation & Classical Normal Theory**
   - $X_n \sim N_{p(n)}(\theta_n, I)$ $p(n) \nearrow$ with $n$ (MVN)

2. **The Threshold Selection Problem**
   - In (MVN) with, say, $\hat{\theta}_i = X_i I\{|X_i| > \hat{t}\}$
   - How to select $\hat{t} = \hat{t}(X)$ “reliably”?

3. **Large Covariance Matrices**
   - $X_n \sim N_{p(n)}(I \otimes \Sigma_{p(n)})$; especially $X_n = \begin{bmatrix} Y_n \\ Z_n \end{bmatrix}$
   - spectral properties of $n^{-1}X_nX_n^T$
   - PCA, CCA, MANOVA