From Stein to minimax predictive density estimation: the sparse normal means case

Iain Johnstone

Statistics, Stanford

Stein Symposium, Singapore, 2019
Agenda

- Stein’s papers on (in)admissible/minimax point estimation
- Predictive density estimation – parallels
- Predictive density estimation – sparse means
  - not fully parallel
Stein on (in)admissible/minimax point estimation

[N.B.: Brown’s 2010 talk (online) has authoritative detail! ]

1955  A necessary and sufficient condition for admissibility  Annals

1956  Inadmissibility of the usual estimator for the mean of a multivariate normal distribution.  3rd Berk. Symp.

1959  Admissibility of Pitman’s estimator of a single location parameter.  Annals

1961  (with James, W.)  Estimation with quadratic loss.  4th Berk. Symp.
Stein on (in)admissible/minimax point estimation

[N.B.: Brown’s 2010 talk (online) has authoritative detail! ]

1955 A necessary and sufficient condition for admissibility Annals
1956 Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. 3rd Berk. Symp.
1959 Admissibility of Pitman’s estimator of a single location parameter. Annals
1961 (with James, W.) Estimation with quadratic loss. 4th Berk. Symp.

“He [Kiefer] told me that Stein was doing some really interesting work on admissibility and I should take a look at that. Statistics was lovely in those days; I essentially had to read five papers to know all the necessary background.” L. Brown, Stat. Sci. (2005)

[The other two: Blackwell (1951), Hodges-Lehmann (1951)]
Stein on (in)admissible/minimax point estimation

1960 Multiple regression   Hotelling vol.

1964 Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean.   Annals ISM


1966 An approach to the recovery of inter-block information in balanced incomplete block designs.   Neyman vol.

1974/81 Estimation of the mean of a multivariate normal distribution.   Prague Symp./ Annals   74 △ 81 ⪞ 74 ∩ 81 !!

1986 Lectures on the theory of estimation of many parameters   Leningrad Seminar
Predictive density estimation
Predictive Estimation

Past data $X$, future data $Y$, $\ X \perp \ Y | \theta$

$$f(y|X = x) \quad \text{predictive density}$$

Books: Aitchison-Dunsmore 75, Geisser 93

Applications: game theory, econometrics, information theory, machine learning, mathematical finance, sports betting

Recent focus: high-dimensional models
Agenda

- (Gaussian) predictive density estimation
  - parallels to quadratic loss
- High dimensional version under sparsity
  - asymptotic minimaxity
- Sparse (i.e. spike & slab) priors, especially
  - uniform slab
  - discrete priors: not parallel
Agenda

- (Gaussian) predictive density estimation
  - parallels to quadratic loss

- High dimensional version under sparsity
  - asymptotic minimaxity

- Sparse (i.e. spike & slab) priors, especially
  - uniform slab
  - discrete priors: not parallel
Predictive Estimation

Idealized version: Observe $X$ from $p(x|\theta) \sim N(\theta, \nu_x I_n)$.

Seek to predict distribution of a independent future observation:
$p(y|\theta) \sim N(\theta, \nu_y I_n)$.

Seek an estimator $\hat{p}(y|x)$ of $p(y|\theta)$. 
Idealized version: Observe $X$ from

$$p(x|\theta) \sim N(\theta, \nu_x l_n).$$

Seek to predict distribution of a independent future observation:

$$p(y|\theta) \sim N(\theta, \nu_y l_n).$$

Seek an estimator $\hat{p}(y|x)$ of $p(y|\theta)$.

E.g. given prior $\pi(d\theta)$ and posterior $\pi(d\theta|x)$, form the Bayes predictive density

$$\hat{p}_\pi(y|x) = \int p(y|\theta)\pi(d\theta|x).$$
Bayes predictive density

\[ \hat{p}_\pi(y|x) = \int p(y|\theta)\pi(d\theta|x) \]

Predictive mean = posterior mean:

\[ \int y\hat{p}_\pi(y|x)dy = \int \theta\pi(d\theta|x) = \hat{\theta}_\pi(x) \]

⇒ intuition from, & parallels with, point estimation.
Bayes predictive density

\[ \hat{p}_\pi(y|x) = \int p(y|\theta)\pi(d\theta|x) \]

Predictive mean = posterior mean:

\[ \int y\hat{p}_\pi(y|x)dy = \int \theta\pi(d\theta|x) = \hat{\theta}_\pi(x) \]

⇒ intuition from, & parallels with, point estimation.

Later: Assume \( \theta = (\theta_i) \) is sparse:

\[ \theta_i \overset{\text{iid}}{\sim} \pi = (1 - \eta)\delta_0 + \eta\nu. \]
Univariate and multivariate versions

Seek \( p_\pi(y|x) \)
Univariate and multivariate versions

Seek \( p_{\pi}(y|x) \)

Seek \( \prod_i p_{\pi}(y_i|x_i) \)
Some predictive density examples

“Plug in” estimate

\[ \hat{p}_{\mu(y|x)} \leftrightarrow N(\hat{\mu}(x), v_y) \quad \hat{\mu}(x) = x \]

- Not Bayes!
Some Bayes predictive density examples

| $\pi(d\mu)$  | $\pi(d\mu|x)$ | $\hat{p}_\pi(y|x)$ |
|-------------|---------------|------------------|
| $\delta_0$  | $\delta_0$    | $N(0, \nu_y)$    |
Some Bayes predictive density examples

| $\pi(d\mu)$ | $\pi(d\mu|x)$ | $\hat{p}_\pi(y|x)$ |
|------------|---------------|------------------|
| $\delta_0$ | $\delta_0$    | $N(0, \nu_y)$   |
| $d\mu$     | $N(x, \nu_x)$ | $N(x, \nu_y + \nu_x)$ |
Some Bayes predictive density examples

\[
\begin{align*}
\pi(d\mu) & \quad \pi(d\mu | x) & \quad \hat{p}_\pi(y | x) \\
\delta_0 & \quad \delta_0 & \quad N(0, v_y) \\
d\mu & \quad N(x, v_x) & \quad N(x, v_y + v_x) \\
N(0, \alpha / (1-\alpha)) & \quad N(\alpha x, \alpha \frac{v_x}{\alpha + (1-\alpha)v_x}) & \quad N(\alpha x, v_y + \alpha v_x)
\end{align*}
\]
Evaluating estimators

Use Kullback-Leibler "distance" from $\hat{p}(y|x)$ to $p(y|\theta)$:

$$L(\theta, \hat{p}(\cdot|x)) = \int p(y|\theta) \log \frac{p(y|\theta)}{\hat{p}(y|x)} dy$$

(Aitchison, 75)

Average over $p(x|\theta)$ to get the Kullback-Leibler (KL) risk:

$$\rho(\theta, \hat{p}) = \int p(x|\theta) L(\theta, \hat{p}(\cdot|x)) dx$$

NB: For $\pi(d\theta)$, the Bayes est. for $\rho$ is the predictive density

$$\hat{p}_\pi(y|x) = \int p(y|\theta) \pi(\theta|x) d\theta.$$
Some parallels with quadratic loss

Mixture representation of K-L risk
Parallels - uniform prior & admissibility

**Quadratic:** \[ X \sim N_n(\theta, v_x I_n) \]
\[ \rho_Q(\theta, \hat{\theta}) = E_\theta ||\hat{\theta} - \theta||^2 \]

\[ \hat{\theta}_{MLE}(x) = x \]  
- formal Bayes for \( \pi_U \)
- minimax
- admissible for \( n = 1, 2 \)

Stein 1956,59; J-Stein 61

**Predictive:** \[ Y \sim N_n(\theta, v_y I_n) \]
\[ \rho(\theta, \hat{\theta}) = \int p(x|\theta)p(y|\theta) \log \frac{p(y|\theta)}{\hat{p}(y|x)} \, dx \, dy \]

\[ \hat{p}_U(y|x) \]  
- formal Bayes for \( \pi_U \)
- minimax
- admissible for \( n = 1, 2 \)

Brown, George, Xu, 2008
Parallels - Bayes rule as a perturbation

**Quadratic:** Marginal density $m_\pi(x)$

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

[Brown, 1971]
Parallels - Bayes rule as a perturbation

**Quadratic:** Marginal density $m_\pi(x)$

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

[Brown, 1971]

**Predictive:**

$$\hat{p}_\pi(y|x) = \hat{p}_U(y|x) \frac{m_\pi(w, v_w)}{m_\pi(x, v_x)}$$

[George, Liang, Xu, 2006]

$$W = \frac{v_y X + v_x Y}{v_x + v_y}, \quad v_w = \frac{v_x v_y}{v_x + v_y} < v_x. \quad \text{oracle variance}$$
Parallels - Stein formulas

**Quadratic:** Stein’s unbiased estimate of risk:

\[
\rho_Q(\theta, \hat{\theta}_{MLE}) - \rho_Q(\theta, \hat{\theta}_\pi) = -4E_\theta \left[ \frac{\nabla^2 \sqrt{m_\pi(X)}}{\sqrt{m_\pi(X)}} \right]
\]

[Stein, 1981; Brown, 1971]

**Predictive:**

\[
\rho(\theta, \hat{\rho}_U) - \rho(\theta, \hat{\rho}_\pi) = \int_{v_w}^{v_x} -2E_{\theta,v} \left[ \frac{\nabla^2 \sqrt{m_\pi(Z, v)}}{\sqrt{m_\pi(Z, v)}} \right] dv
\]

[George, Liang, Xu, 2006]

e.g. \(\pi_H(\theta) = \|\theta\|^{2-n} \Rightarrow \hat{\rho}_{\pi_H}\) dominates \(\hat{\rho}_U\)  

[Komaki, 2001]
Mixture Representation of Kullback-Leibler Risk

Suppose

\[ X|\theta \sim N(\theta, \nu_x I_n), \quad Y|\theta \sim N(\theta, \nu_y I_n) \quad \nu_w = \nu_x \nu_y / (\nu_x + \nu_y) \]

Intermediate variances: If \( Z|\theta \sim N(\theta, \nu I_n) \), let

\[ \rho^\nu_{\theta Q}(\theta, \hat{\theta}_\pi) = E_{\theta, \nu} ||\hat{\theta}_\pi(Z) - \theta||_2^2, \quad \hat{\theta}_\pi(Z) = E(\theta|Z) \]

**Proposition** *(Brown, George, Xu, 2008)*

\[ \rho(\theta, \hat{\theta}_\pi) = \int_{\nu_w}^{\nu_x} \rho^\nu_{\theta Q}(\theta, \hat{\theta}_\pi) \frac{d\nu}{2\nu^2}. \]

Predictive risk = weighted ave of quadratic risk, \( \forall \in (\nu_w, \nu_x) \).
High dimensional version under sparsity
Consider product predictive density:

\[ \hat{p}(y|x) = \prod_{i=1}^{n} \hat{p}_i(y_i|x_i) \]

KL risk decomposes

\[ \rho(\theta, \hat{p}) = \sum_{i=1}^{n} \rho(\theta_i, \hat{p}_i) \]

Consider sparse \( \theta \), with at most \( s \) non-zero entries:

\[ \theta \in \Theta_n[s] = \{ \theta \in \mathbb{R}^n : ||\theta||_0 \leq s \} \]
Example: $X$ and $Y$ result from a *sparsifying transform* – e.g. wavelet transform of phone call center data (Brown et. al. 2005)

$X_p(t) =$ volume of calls at time $t$, on *past* day(s)
$\theta(t) =$ mean volume of calls at time $t$ on typical days
$Y_f(t) =$ volume of calls at time $t$, on a *future* day
Example: $X$ and $Y$ result from a *sparsifying transform* – e.g. wavelet transform of phone call center data (Brown et. al. 2005)

$$X_p(t) = \text{volume of calls at time } t, \text{ on past day(s)}$$
$$\theta(t) = \text{mean volume of calls at time } t \text{ on typical days}$$
$$Y_f(t) = \text{volume of calls at time } t, \text{ on a future day}$$

In wavelet transform domain: $(\theta_i)$ will be sparse:

$$X(t) \rightarrow x = (x_i, i = 1, \ldots, n)$$
$$\theta(t) \rightarrow \theta = (\theta_i, i = 1, \ldots, n)$$
$$Y(t) \rightarrow y = (y_i, i = 1, \ldots, n)$$

What can we say about KL risk under sparsity assumption:

$$\inf_{\hat{\rho}} \sup_{\theta \in \Theta_n[s]} R_{KL}(\theta, \hat{\rho})$$
Multivariate and Univariate

\[ \hat{p}(y|x) = \prod_{i} \hat{p}(y_i|x_i) \]

For sparse \( \theta \in \Theta_n[s_n] \):

\[ \rho(\theta, \hat{p}) = (n - s_n)\rho(0, \hat{p}) + \sum_{\theta_i \neq 0} \rho(\theta_i, \hat{p}) \]
Multivariate and Univariate

\[ \hat{p}(y|x) = \prod_i \hat{p}(y_i|x_i) \]

For sparse \( \theta \in \Theta_n[s_n] \):

\[ \rho(\theta, \hat{p}) = (n - s_n)\rho(0, \hat{p}) + \sum_{\theta_i \neq 0} \rho(\theta_i, \hat{p}) \]

Role of maximum univariate risk: With \( \eta_n = s_n/n \),

\[ \sup_{\Theta_n[s_n]} \rho(\theta, \hat{p}) = n(1 - \eta_n)\rho(0, \hat{p}) + s_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}) \]

\[ \sim n\eta_n \sup_{\theta \in \mathbb{R}} \rho(\theta, \hat{p}) \]
Asymptotic Sparse Minimax risk

Let \( X \sim N(\theta, I_n) \perp \perp Y \sim N(\theta, rI_n), \quad [v_x = 1, \quad r = v_y = \frac{v_y}{v_x}] \)

\[ R_N(\Theta_n[s_n]) = \inf_{\hat{\rho}} \sup_{\theta \in \Theta_n[s_n]} \rho(\theta, \hat{\rho}). \]

\(^1\)Mukherjee-J, 2015
Asymptotic Sparse Minimax risk

Let \( X \sim N(\theta, I_n) \perp \perp Y \sim N(\theta, rI_n), \quad [v_x = 1, \ r = v_y = \frac{v_y}{v_x}] \)

\[
\begin{align*}
R_N(\Theta_n[s_n]) &= \inf \hat{\rho} \sup_{\theta \in \Theta_n[s_n]} \rho(\theta, \hat{\theta}).
\end{align*}
\]

**Theorem**\(^1\) Fix \( r \in (0, \infty) \). If \( \eta_n = s_n/n \to 0 \), then

\[
R_N(\Theta_n[s_n]) \sim \frac{1}{1 + r} s_n \log(n/s_n) \sim n\eta_n \frac{\lambda_f^2}{2r},
\]

with predictive scale \( \lambda_f^2 = 2v_w \log \eta_n^{-1} \). \([v_w^{-1} = 1 + r^{-1}]\)

\(^1\)Mukherjee-J, 2015
Asymptotic Sparse Minimax risk

Let \( X \sim N(\theta, I_n) \perp Y \sim N(\theta, rI_n) \), \([v_x = 1, \; r = v_y = \frac{v_y}{v_x}]\)

\[
R_N(\Theta_n[s_n]) = \inf_{\hat{p}} \sup_{\theta \in \Theta_n[s_n]} \rho(\theta, \hat{p}).
\]

**Theorem**
Fix \( r \in (0, \infty) \). If \( \eta_n = s_n/n \to 0 \), then

\[
R_N(\Theta_n[s_n]) \sim \frac{1}{1 + r} s_n \log(n/s_n) \sim n\eta_n \frac{\lambda_f^2}{2r},
\]

with predictive scale \( \lambda_f^2 = 2v_w \log \eta_n^{-1} \).

**Remark.** For quadratic loss, with estimation scale \( \lambda_e^2 = 2 \log \eta_n^{-1} \),

\[
R_{N,Q}(\Theta_n[s_n]) \sim 2s_n \log(n/s_n) \sim n\eta_n \lambda_e^2,
\]

\[^1\text{Mukherjee-J, 2015}\]
Understanding $\lambda_e^2 = 2 \log \eta_n^{-1} > \lambda_f^2 = 2 \nu_w \log \eta_n^{-1}$

For $X \sim N(\theta, \nu)$, $0 \leq \theta \leq \sqrt{\nu} \lambda_e$ are ‘invisible’ for point est’n:

$$\pi = (1 - \eta) \delta_0 + \eta \delta_{\sqrt{\nu} \lambda_e} \quad \Rightarrow \quad \hat{\theta}_\pi \approx 0 \quad \text{for} \quad x \sim N(\theta, 1)$$

For $\nu_w \leq \nu \leq 1$, the “common invisibility region” is

$$0 \leq \theta \leq \sqrt{\nu_w} \lambda_e =: \lambda_f$$

With $\pi[\eta, \lambda_f] = (1 - \eta) \delta_0 + \eta \delta_{\lambda_f}$ and $q(\lambda_f; \hat{\theta}_\pi, \nu; \nu) \gtrapprox \lambda_f^2$,

$$\rho(\lambda_f, \hat{\rho}_\pi) = \int_{\nu_w}^{1} q(\lambda_f; \hat{\theta}_\pi, \nu; \nu) \frac{dv}{2\nu^2} \gtrapprox \lambda_f^2 \int_{\nu_w}^{1} \frac{dv}{2\nu^2} = \frac{\lambda_f^2}{2r}$$
Looking for minimax Bayes rules
Background

Mukherjee-J (2015) used a threshold rule

\[
\hat{p}_T(y_i|x_i) = \begin{cases} 
\hat{p}_U(y_i|x_i) & \text{if } |x_i| > v^{-1/2} \lambda \\
\hat{p}_{CL}(y_i|x_i) & \text{if } |x_i| \leq v^{-1/2} \lambda
\end{cases}
\]

to prove asymptotic minimaxity – discontinuous!

This talk: proper Bayes predictive density estimates that are asymptotically minimax

Key tool: predictive risk identity based on decomposition

\[
p_\pi(y|x) = \phi(y|0, r) \frac{N(x, y)}{D(x)}
\]
Predictive risk identity for sparse priors

**Proposition:** For a sparse prior \( \pi = (1 - \eta)\delta_0 + \eta\nu \)

\[
\rho(\theta, \hat{p}_\pi) = \rho(\theta, \hat{p}_{\delta_0}) - \mathbb{E}\log N_{\theta,\nu}(Z) + \mathbb{E}\log D_\theta(Z),
\]

\[
= \frac{\theta^2}{2r} - \mathbb{E}\log N_{\theta,\nu}(Z) + \mathbb{E}\log D_\theta(Z)
\]

where \( D_\theta(Z) = N_{\theta,1}(Z) \), and

\[
N_{\theta,\nu}(Z) = 1 + \frac{\eta}{1 - \eta} \int \exp\left\{ \frac{\mu}{\sqrt{\nu}} \left( Z + \frac{\theta}{\sqrt{\nu}} \right) - \frac{\mu^2}{2\nu} \right\} \nu(d\mu).
\]
Predictive risk identity for sparse priors

**Proposition:** For a sparse prior \( \pi = (1 - \eta)\delta_0 + \eta \nu \)

\[
\rho(\theta, \hat{p}_\pi) = \rho(\theta, \hat{p}_\delta) - \mathbb{E} \log N_{\theta, \nu}(Z) + \mathbb{E} \log D_{\theta}(Z),
\]

\[
= \frac{\theta^2}{2r} - \mathbb{E} \log N_{\theta, \nu}(Z) + \mathbb{E} \log D_{\theta}(Z)
\]

where \( D_{\theta}(Z) = N_{\theta, 1}(Z) \), and

\[
N_{\theta, \nu}(Z) = 1 + \frac{\eta}{1 - \eta} \int \exp \left\{ \frac{\mu}{\sqrt{\nu}} \left( Z + \frac{\theta}{\sqrt{\nu}} \right) - \frac{\mu^2}{2\nu} \right\} \nu(d\mu).
\]

Use with

- \( \nu = \text{uniform slab} \quad (2l)^{-1}I\{|\theta| \leq l\} \rightarrow \hat{p}_S[l] \)

- \( \nu = \text{discrete prior} \quad (\text{e.g.} \quad \frac{1}{2}\delta_\lambda + \frac{1}{2}\delta_{-\lambda}, \text{countably discrete}) \)
Spike and Uniform slab - asymptotic minimaxity

\[ \theta \to \rho(\theta, \hat{\rho}_S[l]) \] is unbounded for \( \theta \) large, so set

\[ \Theta_n[s, t] = \{ \theta \in \mathbb{R}^n : \|\theta\|_0 \leq s; \|\theta\|_\infty \leq t \} \]

**Theorem** If diameter \( t_n \) satisfies

\[ \lambda_f = \lambda_n \ll t_n, \quad \log t_n \ll \frac{1}{2} \lambda_n^2, \]

then \( \hat{\rho}_S[t_n] \) is asy. minimax:

\[ \max_{\Theta_n[s_n, t_n]} \rho(\theta, \hat{\rho}_S[t_n]) \sim R_N(\Theta_n[s_n, t_n]) \sim R_N(\Theta_n[s_n]) \]

Fails for very large \( t_n \): if \( \log t_n \sim \frac{\beta}{2} \lambda_n^2 \),

\[ \min_l \max_{\Theta_n[s_n, t_n]} \rho(\theta, \hat{\rho}_S[l]) \geq \eta_n^2 (1+\beta) \frac{\lambda_n^2}{2r} \]
(Infinite) Discrete Priors
Least favorable sparse prior

Least favorable prior maximizes \( \sup \{ B(\pi) : \pi = (1 - \eta)\delta_0 + \eta\nu \} \)

General minimax theory yields – for fixed \( \eta \) – :

- a **unique** least favorable prior \( \pi_\eta \)
- \( \pi_\eta \) is symmetric, countable discrete
- **maxima of** \( \rho(\theta, \hat{\pi}_\eta) \subset \) **support of** \( \pi_\eta \)
Understanding $\pi_\eta$

For fixed $\eta$, complicated numerical optimization needed:

Goal: find *simply described* discrete priors

yielding *asymptotically* minimax priors

Look for *sparse* spacing

- leads to periodic structure (for large $\theta$)
- easier analysis (conceptually)
Sparse discrete priors ctd

Prior: \( \pi_j \) at location \( \mu_j \)

Posterior: \( \pi_j^x \propto \pi_j \exp\{-\frac{1}{2}(x - \mu_j)^2\} \)
Sparse discrete priors ctd

Prior: \( \pi_j \) at location \( \mu_j \)

Posterior ratio: \[
\frac{\pi_j^x}{\pi_{j-1}^x} \propto \frac{\pi_j}{\pi_{j-1}} \exp\left\{-\frac{1}{2}(x - \mu_j)^2 + \frac{1}{2}(x - \mu_{j-1})^2\right\}
\]

For large spacings \( \mu_j - \mu_{j-1} = O(\lambda) \),

▷ use a two-point posterior to approximate \( \rho(\theta, \hat{\pi}) \)
Discrete posteriors and Australian football
Discrete posteriors and Australian football
Spacing of first atom?
Spacing of first atom on scale $\lambda$

$$\pi_a = (1 - \eta)\delta_0 + \eta\delta_{a\lambda} \quad [\lambda = \lambda_f \text{ from now on, predictive scale}]$$

What value of $a$ is asymptotically least favorable?

$$B(\pi_a) = (1 - \eta)\rho(0, \hat{p}_{\pi_a}) + \eta\rho(a\lambda, \hat{p}_{\pi_a})$$

$$\sim \frac{\eta\lambda^2}{2r} \sigma(a) \leq \frac{\eta\lambda^2}{2r} \quad (\text{as } \eta \to 0)$$

equality only if $a = 1$, i.e. predictive scale spacing.
A uniform grid prior fails for small $r$

Linearly spaced atoms with geometric decay of prior weights:

$$\mu_j = j \lambda, \quad \pi_j = c(\eta) \eta^{j\nu},$$
A uniform grid prior fails for small $r$

Linearly spaced atoms with geometric decay of prior weights:

$$\mu_j = j\lambda, \quad \pi_j = c(\eta)\eta^j, \quad \rho(\alpha\lambda, \hat{p}_G) \sim \frac{\lambda^2}{2r}\sigma(\alpha) + O(\lambda)$$
A uniform grid prior fails for small $r$

Linearly spaced atoms with geometric decay of prior weights:

\[ \mu_j = j\lambda, \quad \pi_j = c(\eta)\eta^{j\nu}, \quad \rho(\alpha\lambda, \hat{\rho}_G) \sim \frac{\lambda^2}{2r}\sigma(\alpha) + O(\lambda) \]

\[ \max_{\lambda \leq \theta \leq 2\lambda} \sigma(\theta) = \max\{1, 1 + \gamma_r(\gamma_r - 2r)\} \geq 1 \]

\[ \Rightarrow \hat{\rho}_G \text{ is asy. minimax if and only if } r \geq r_0 = (\sqrt{5} - 1)/4. \]
Contrast with point estimation

Equispaced geometric prior is asymptotically least favorable, (to leading order): (J, 1994)

\[ \mu_j = j \lambda e, \quad \pi_j = c(\eta) \eta^j, \quad \lambda = \sqrt{2 \log \eta^{-1}} \]

⇒ The point/predictive estimation parallel fails here when

\[ r = \frac{V_y}{V_x} < (\sqrt{5} - 1)/4 \approx 0.3090. \]
Asy. minimaxity for $r < (\sqrt{5} - 1)/4$
Bi-grid prior

Reduce spacing for ‘inner’ intervals: for \( b < 1 \):

\[
\mu_j = \alpha_j \lambda \\
\pi_j = c(\eta) \eta \eta^{\beta_j-1}
\]

\[
\alpha_j = \begin{cases} 
1 + b(j - 1) & 1 \leq j \leq K \\
\alpha_K + j - K & j > K 
\end{cases}
\]

\[
\beta_j = \begin{cases} 
1 + b^2(j - 1) & 1 \leq j \leq K \\
\beta_K + j - K & j > K 
\end{cases}
\]

\[
K \geq 1 + \lceil 2b^{-3/2} \rceil
\]
Bi-grid risk

Inner zone: decrease $b$ to achieve minimaxity (outer zone is ok):

$$\max_{\theta \in \text{inner}} \rho(\theta, \hat{p}_B) = \max\{1, 1 + \gamma_r b(\gamma_r b - 2r)\}$$

$$\max_{\theta \in \text{outer}} \rho(\theta, \hat{p}_B) = \gamma_r^2 < 1.$$
Theorem Assume that

\[ b \leq \min \left\{ \frac{4r(1 + r)}{1 + 2r}, 1 \right\}, \quad K \geq 1 + \lceil 2b^{-3/2} \rceil, \]

Then \( \hat{p}_B \) is asymptotically minimax: if \( s_n/n \to 0 \), then

\[
\max_{\Theta_n[s_n]} \rho(\theta, \hat{p}_B) = R_N(\Theta_n[s_n])(1 + o(1))
\]
Conclusion

- asymptotically minimax rules from sparse priors
- exception to point/predictive parallel; arXiv: 1707.04380
- future: fuller study of discrete priors
Conclusion

- asymptotically minimax rules from sparse priors
- exception to point/predictive parallel; arXiv: 1707.04380
- future: fuller study of discrete priors

G. Elving: (early 1970s)

“... after I met Stein I wondered why anyone else did mathematical statistics.”
Conclusion

- asymptotically minimax rules from sparse priors
- exception to point/predictive parallel; arXiv: 1707.04380
- future: fuller study of discrete priors

A reply to G. Elving:

“Charles mind has always been loaded with insights and ideas. He has been generous about sharing these in person and at each opportunity in his papers.

Many of these have been pursued and utilized by others. Some raise still vibrant questions that should yet be pursued further.”

L. Brown, 2010
Backup
Understanding

\[ \lambda_e^2 = 2 \log \eta_n^{-1} \]

\[ > \lambda_f^2 = 2v \log \eta_n^{-1} \]
Estimation: Least favorable priors and ‘invisibility’

Two point sparse prior, \( \lambda = \sqrt{2 \log \eta^{-1}}, \quad \mu = \lambda - \sqrt{\lambda} \)
Estimation: Least favorable priors and ‘invisibility’

Two point sparse prior, \( \lambda = \sqrt{2 \log \eta^{-1}} \), \( \mu = \lambda - \sqrt{\lambda} \)

\[
\eta(x) = \pi(\{\mu\} | x = \mu + z) \approx e^{-\lambda(\sqrt{\lambda} - z)} \quad \text{small!}
\]

\[\Rightarrow \delta_\pi(x) = \mathbb{E}[\theta | x] \approx 0 \quad \text{even if} \quad x \sim N(\mu, 1)\]
Estimation: Least favorable priors and ‘invisibility’

Two point sparse prior, \( \lambda = \sqrt{2 \log \eta^{-1}}, \quad \mu = \lambda - \sqrt{\lambda} \)

\[
\eta(x) = \pi(\{\mu\}|x = \mu + z) \approx e^{-\lambda(\sqrt{\lambda} - z)} \quad \text{small!}
\]

\[
\Rightarrow \delta_\pi(x) = \mathbb{E}[\theta|x] \approx 0 \quad \text{even if} \quad x \sim N(\mu, 1)
\]

Thus, for \( X \sim N(\theta, 1), \)

\[
0 \leq \theta \leq \sqrt{2 \log \eta^{-1}} = \lambda_e \quad \text{are ‘invisible’}
\]
Prediction: Invisibility across scales

Rescale: for $X \sim N(\theta, \nu)$,

$$0 \leq \theta \leq \sqrt{\nu} \lambda_e$$

are ‘invisible’

Recall $\nu_w = (1 + r^{-1})^{-1} < 1$, and

$$\rho(\theta, \hat{p}_\pi) = \int_{\nu_w}^{1} q(\theta; \hat{\theta}_\pi, \nu) \frac{d\nu}{2\nu^2}$$

For $\nu_w \leq \nu \leq 1$, the “common invisibility region” is the intersection

$$0 \leq \theta \leq \sqrt{\nu_w} \lambda_e =: \lambda_f$$
Plausible LF prior and minimax risk

\[ \pi[\eta, \lambda_f] = (1 - \eta)\delta_0 + \eta\delta_{\lambda_f} \]

Due to invisibility, as \( \eta \to 0 \), for each \( \nu \),

\[ q(\lambda_f; \hat{\theta}_{\pi, \nu}; \nu) \gtrsim \lambda_f^2 \]

For the predictive risk

\[ \rho(\lambda_f, \hat{p}_{\pi}) \gtrsim \lambda_f^2 \int_{V_w}^1 \frac{d\nu}{2\nu^2} = \frac{\lambda_f^2}{2r} \]

Suggests (asymptotic, univariate) minimax risk \( \sim \frac{\lambda_f^2}{2r} \)
Spike and Uniform slab – unbounded risk

So, ... Change the problem: → bounded $\Theta$:

$$\Theta_n[s, t] = \{\theta \in \mathbb{R}^n : \|\theta\|_0 \leq s; \|\theta\|_\infty \leq t\}$$

Same minimax risk if $t = t_n > \lambda_n$:

$$R_N(\Theta_n[s_n, t_n]) \sim R_N(\Theta_n[s_n])$$
Theorem If diameter $t_n$ satisfies

$$\lambda_f = \lambda_n \ll t_n, \quad \log t_n \ll \frac{1}{2} \lambda_n^2,$$

then $\hat{\rho}_S[t_n]$ is asy. minimax:

$$\max_{\Theta_n[s_n, t_n]} \rho(\theta, \hat{\rho}_S[t_n]) \sim R_N(\Theta_n[s_n, t_n])$$

Fails for very large $t_n$: if

$$\log t_n \sim \frac{\beta}{2} \lambda_n^2,$$

$$\min_l \max_{\Theta_n[s_n, t_n]} \rho(\theta, \hat{\rho}_S[l]) \geq n \eta_n^2 (1+\beta) \frac{\lambda_n^2}{2r}$$
Spike and Uniform slab – risk bounds

Risk bounds OK if \( \log l \ll \lambda_n^2 \), so

- If \( \log t_n \ll \lambda_n^2 \), set \( l = t_n \)
- But, if \( \log t_n \sim \beta \lambda_n^2 \), then no \( l \) works:
  - if \( \log l \geq \beta \lambda_n^2 \), max at \( \theta = \sqrt{1 + \beta \lambda_n} \) is too big
  - if \( \log l < \beta \lambda_n^2 \), max at \( \theta = t_n \) is too big
(Infinite) Discrete Priors