6th BS/IMSC

UNIVERSITAT DE BARCELONA

IMUB Institut de Matemàtica

ims

The Threshold Selection Problem

Given
- data $X_i \sim N(\mu_i, 1), \quad i = 1, \ldots, n$
- some threshold rule, e.g.

$$\hat{\theta}_i = \begin{cases} X_i & |X_i| \geq t \\ 0 & |X_i| < t \end{cases}$$

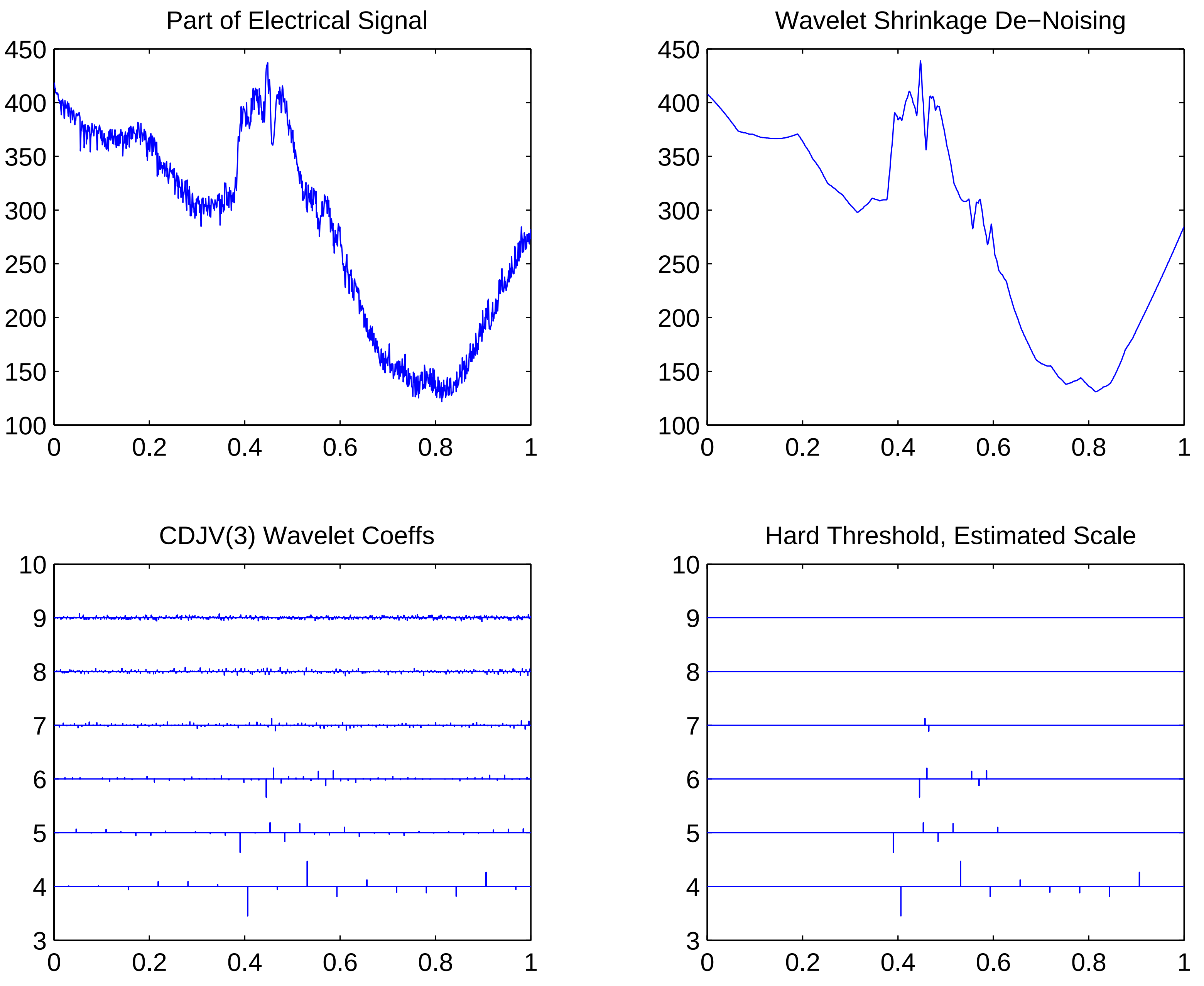
How to "reliably" choose $\hat{t} = \hat{t}(X_1, \ldots, X_n)$?

- case study using Growing Gaussian Models
- well studied, but complicated by **phase changes** (v. sparse $\longrightarrow$ sparse $\longrightarrow$ dense signals μ)
- no 'best' solution;
- An apparently *ad hoc* Empirical Bayes compromise

Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

An “Easy” Example



Transform Shrinkage Approach

$$
\begin{tikzcd}
(y_i) \arrow[d, "W"'] & & \left(\hat{f}(t_i)\right) \\
((d_{jk})) \arrow[rr, "\eta"] & & ((\hat{d}_{jk})) \arrow[u, "W^{-1}"']
\end{tikzcd}
$$

Processing step:

- Operations on *groups* (blocks) $d_j = (d_{jk} : k = 1, \ldots, n_j)$
- $\eta : d_j \longrightarrow \hat{d}_j$

Features of block processing:

- creates more homogeneous subgroups
- ignores cross block dependence

Block structure in many transform problems

Direct data: $y = f + z$

- 1-d signals:
  - might choose subblocks within levels
  - blocking of Fourier coefficients
- Images:
  - wavelet bases: resolution level $\times$ channel
  - ridgelet bases (or frames), brushlets, curvelets ...

Indirect data: $y = Kf + z$

- K (linear) operator: integration, Radon transform, convolution/blurring. . . .
- If $\exists$ near diagonalization in some transform domain

$$y_{Jk} = \alpha_J \theta_{Jk} + z_{Jk}, \qquad k \in B_J, J \in \mathcal{J}$$

Indirect data Examples

- Fourier transform (deconvolution)
- Singular value decomposition (SVD)
- Wavelet-vaguelette decomposition (WVD): $\{v_{Jk}\}, \{w_{Jk}\}$ such that
$$Kv_{Jk} = \alpha_J w_{Jk}$$
- Certain wavelet packet bases: e.g. mirror wavelets for deconvolution

  - Mallat CNES example

‘Truth’ and Blurred Data

Wiener filter vs. Mirror Wavelets

$SNR = 32.7db$

$SNR = 34.1db$

Choice of basis/transform

- Try to choose transform domain so that
  a) signal coefficients are sparse,
  b) noise is relatively white *within blocks*
- "sparse": energy is concentrated in a few compononents

Examples:

Smooth signals:	Fourier basis, energy in low frequencies
Oscillatory signals:	wavelet or cosine packet bases
Point singularities:	wavelet bases
Line singularities:	ridgelet, curvelet bases

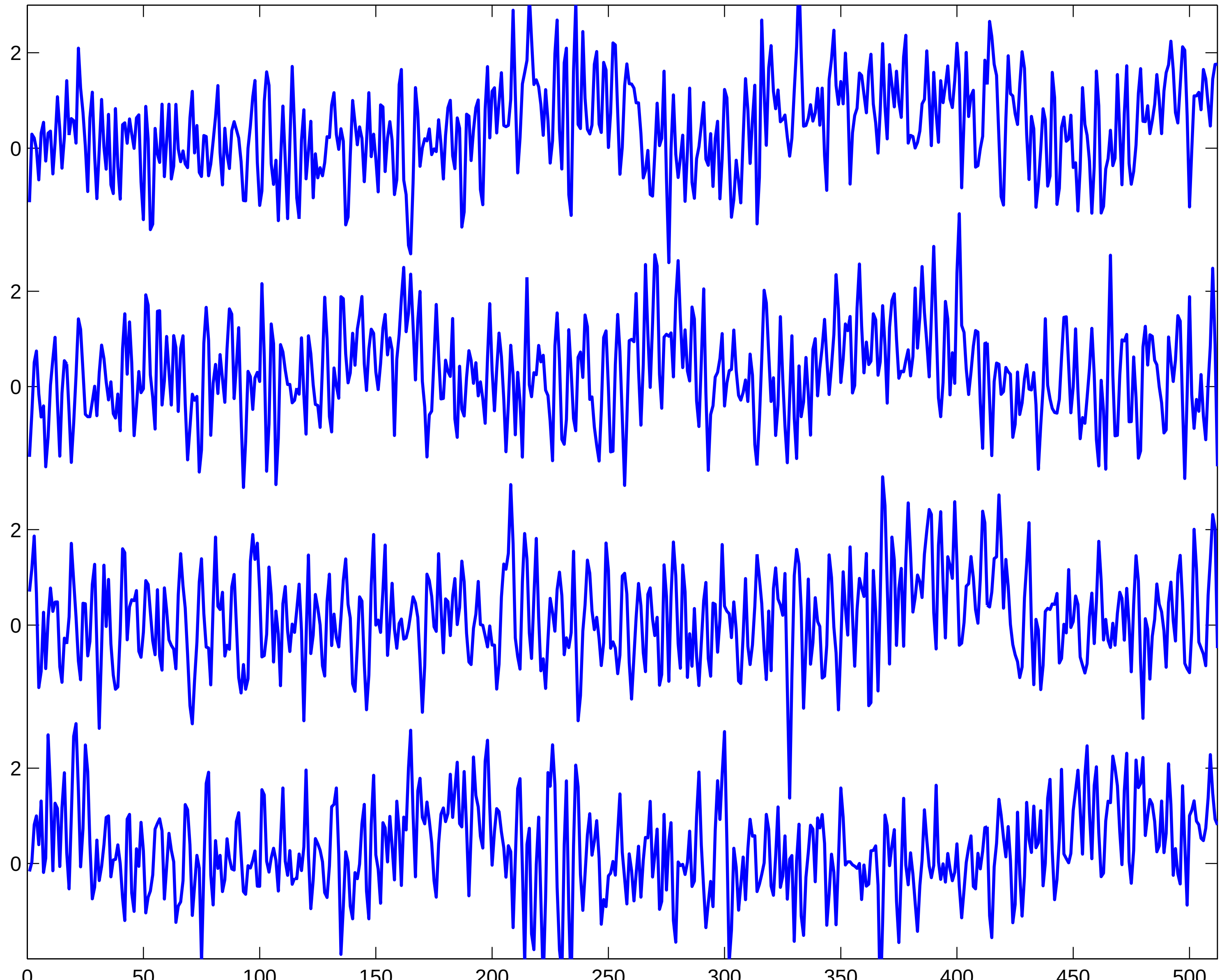
Henceforth, we assume this has been done ...

Agenda

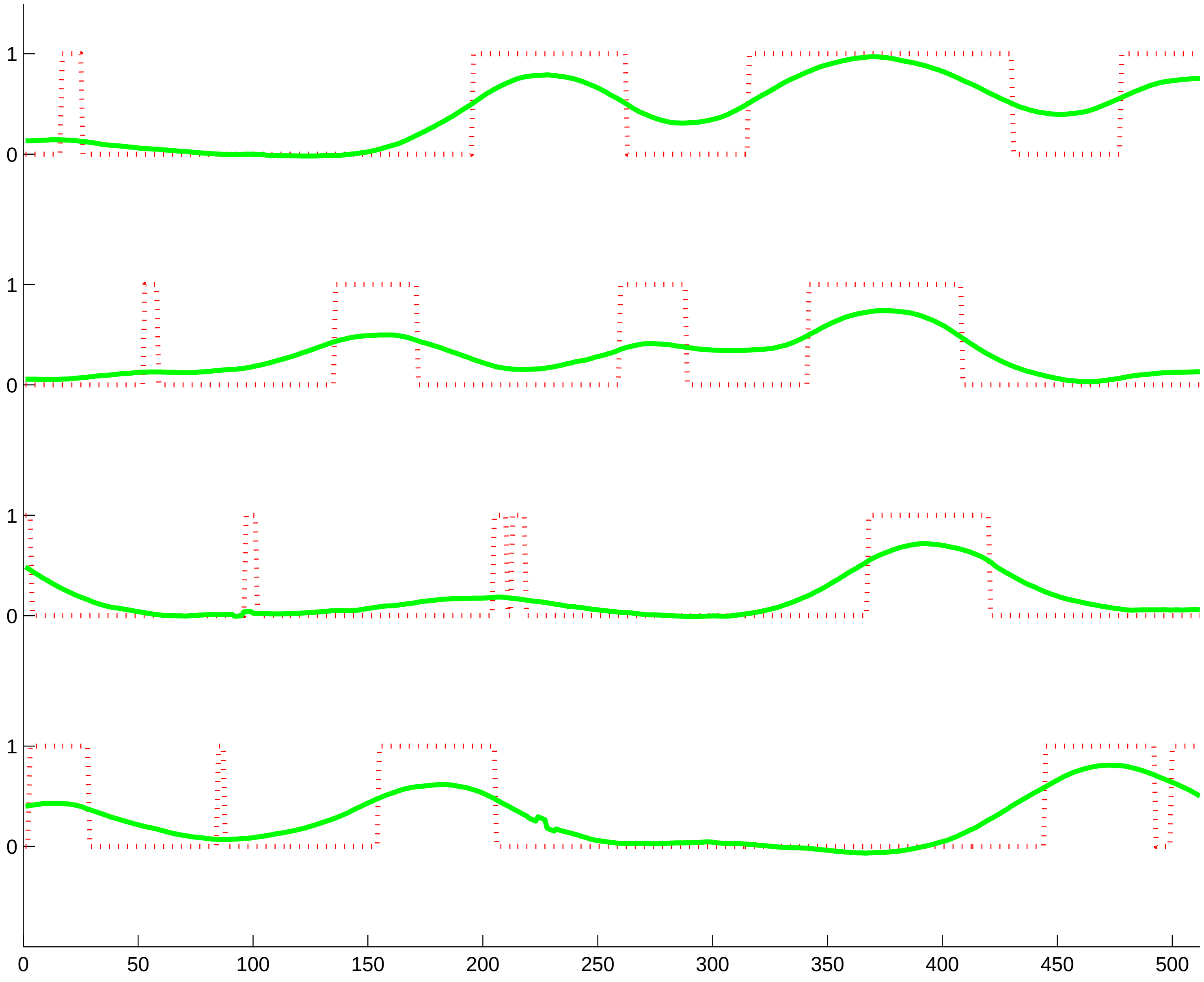
- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Why threshold choice matters, I

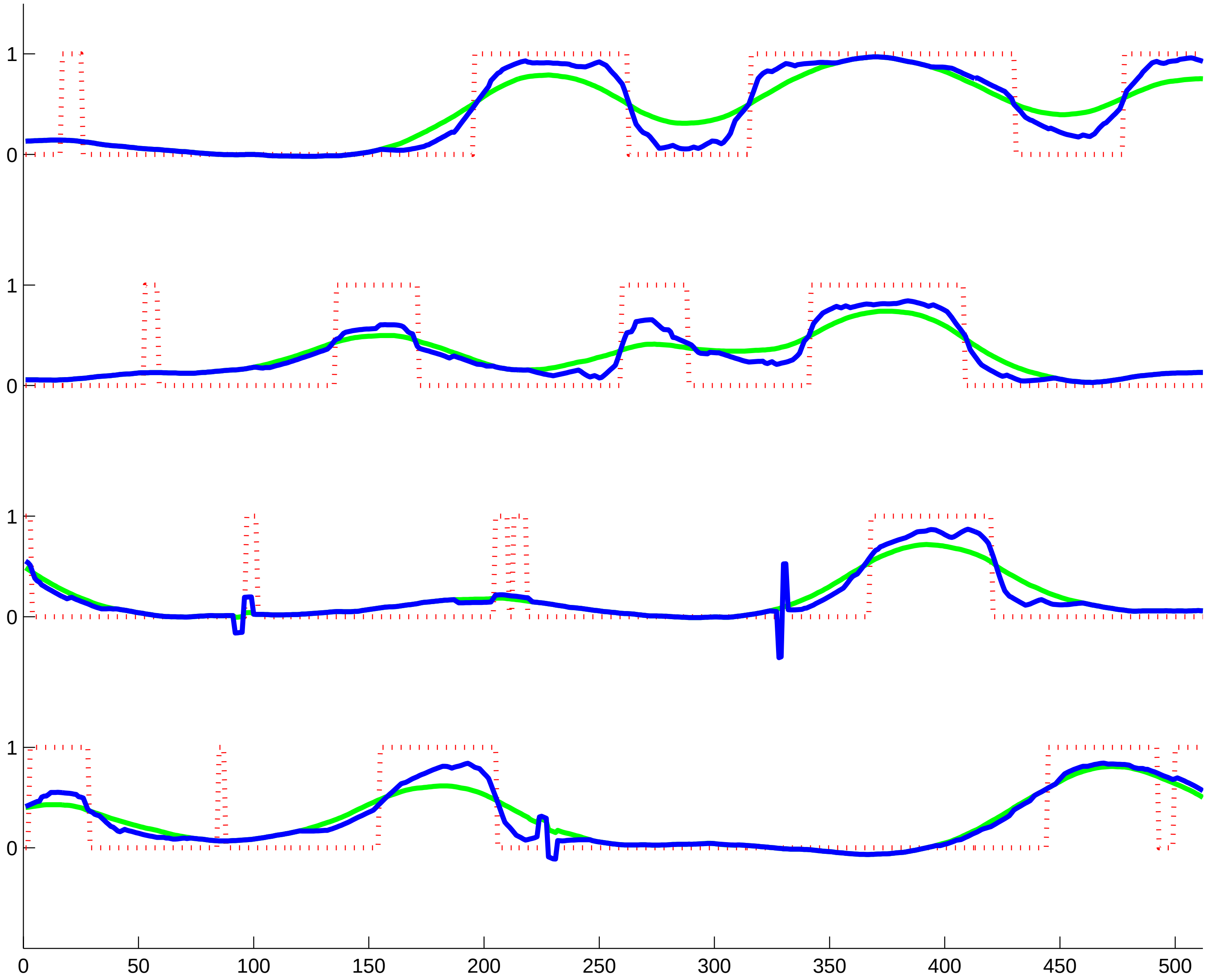
Portion of Ion Channel Signal



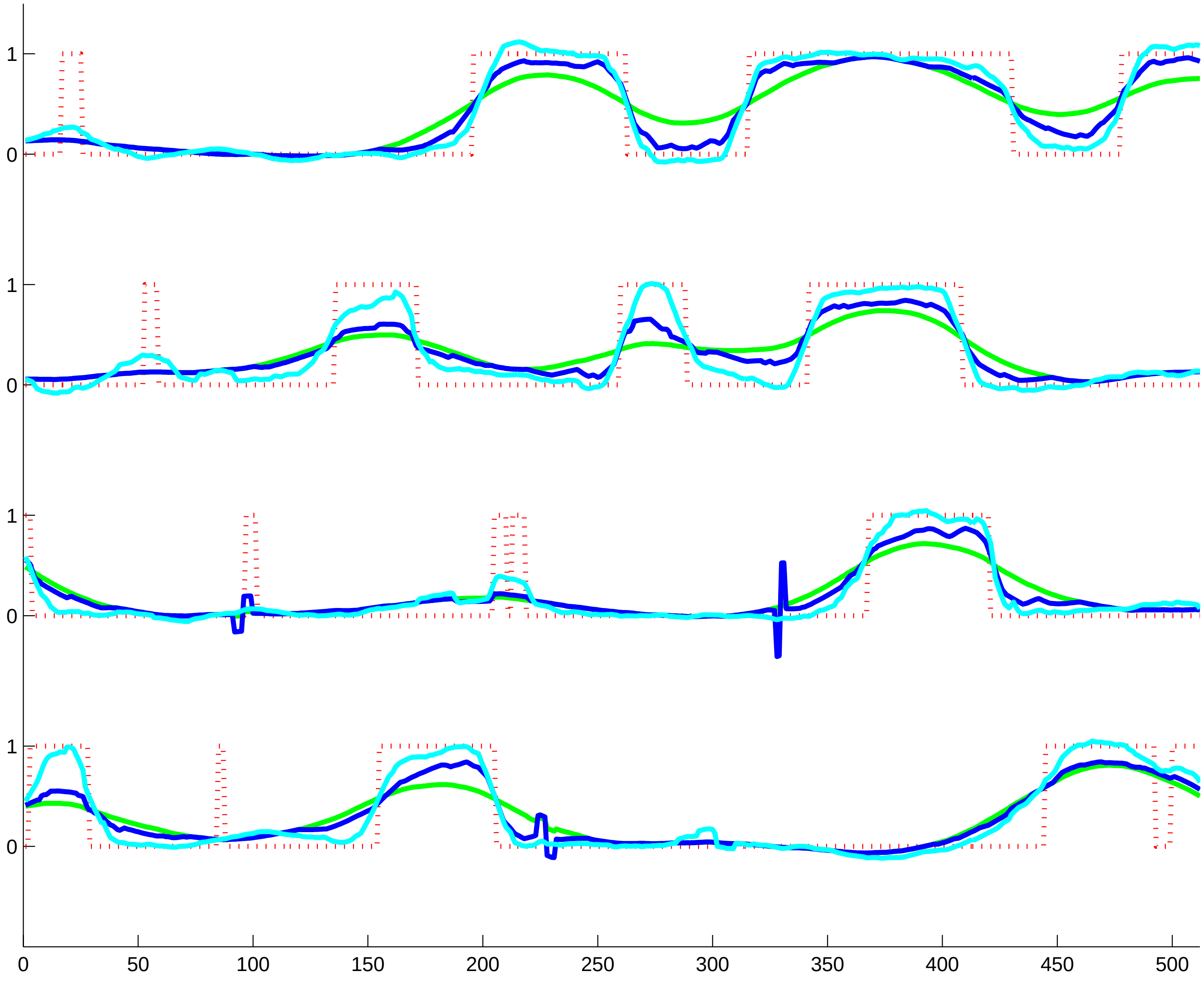
$\sqrt{2\log n}$ thresholds: 8.9 % errors



FDR, $q = 0.05$ thresholds: 6.7% errors



E-Bayes thresholds: 2.9% errors



Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Single sequence problem

Single block (e.g. level in wavelet transform)

$$X_i = \mu_i + z_i, \qquad z_i \overset{ind}{\sim} N(0,1) \quad i = 1, 2, \dots, n$$

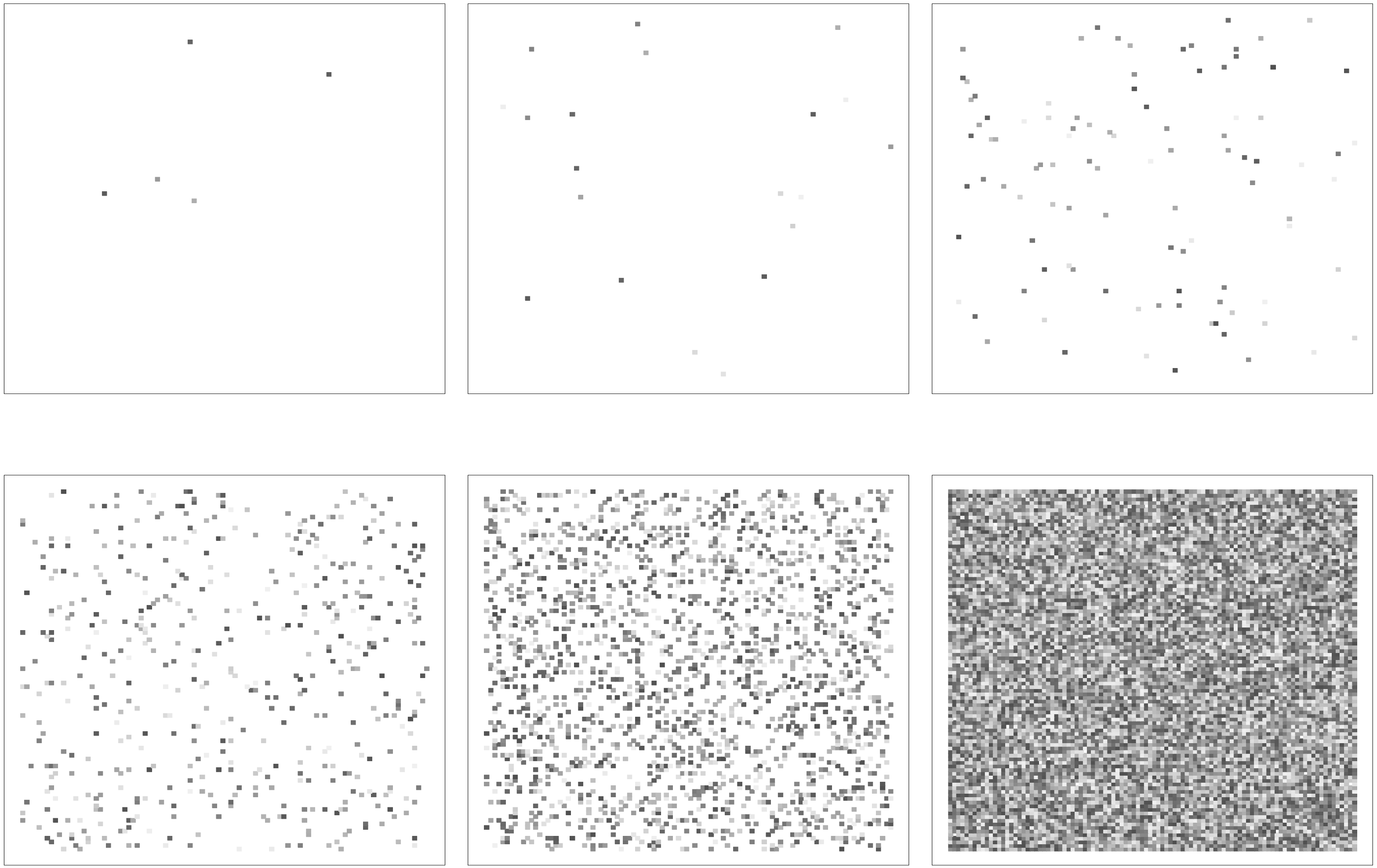
μ is **fixed** & unknown, z is random. ("many normal means problem")

- e.g. for wavelet blocks, $n = 2^j$, but not necessarily
- Want to adapt well to sparsity in the sequence μ,
  while also dealing well with 'dense' cases.

- a "growing Gaussian" model, since $n = \#$ variables $\nearrow$

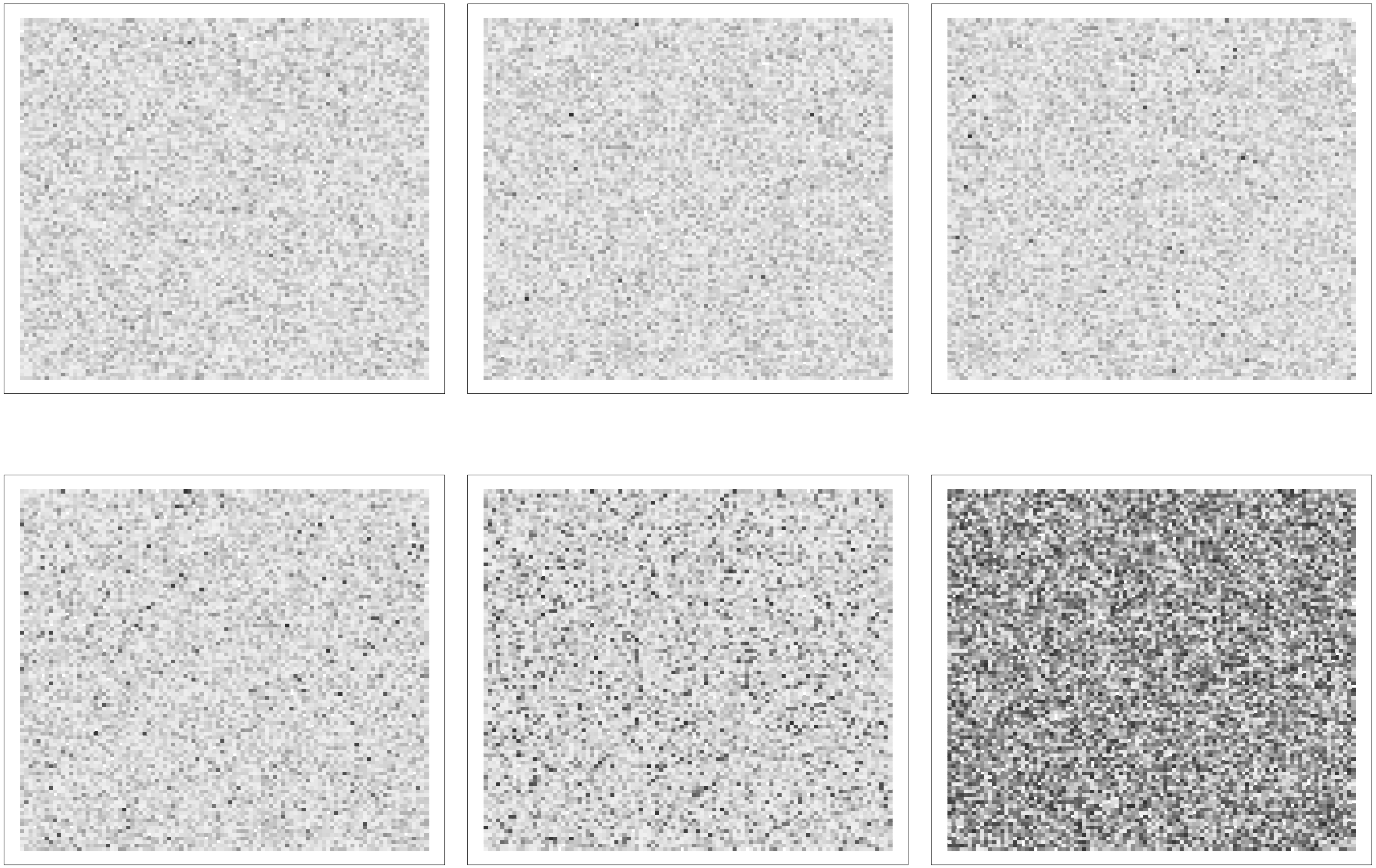
Why threshold choice matters, II

$\mu_i \sim U(-5, 5)$ with probability $.0005, .002, .01, .05, .2, 1$

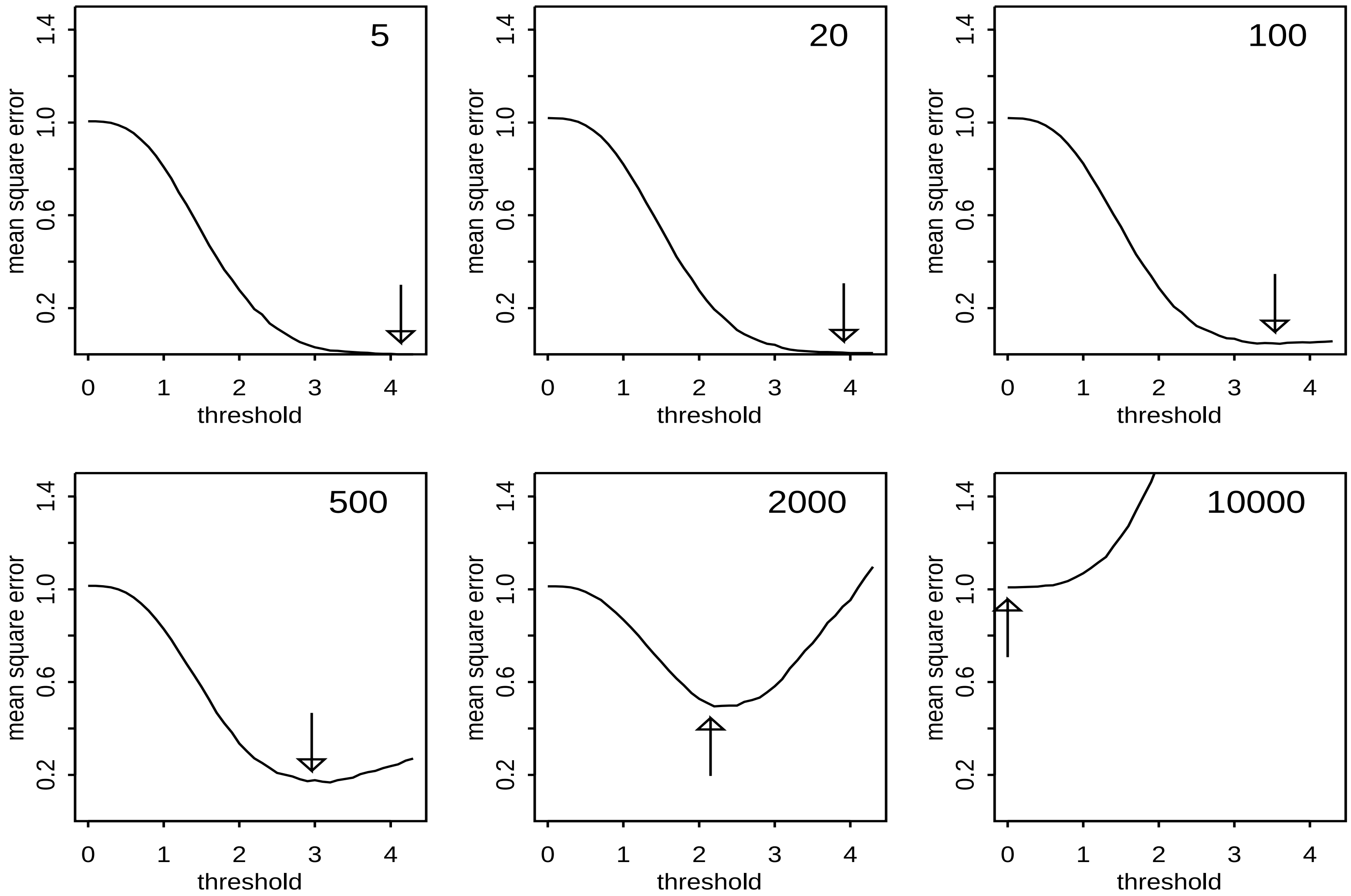


Noise added

$$x_i = \mu_i + z_i, \qquad i = 1, \ldots, 10,000$$



Optimal thresholds vary with sparsity:



Coordinatewise Non-linearity

$$\hat{\mu}_i(x) = \eta(x_i, \hat{t})$$

Extreme Examples: *Hard*: $\eta(x_i, t) = x_i I\{|x_i| > t\}$

Soft: $\eta(x_i, t) = \text{sign}(x)(|x_i| - t)_+$

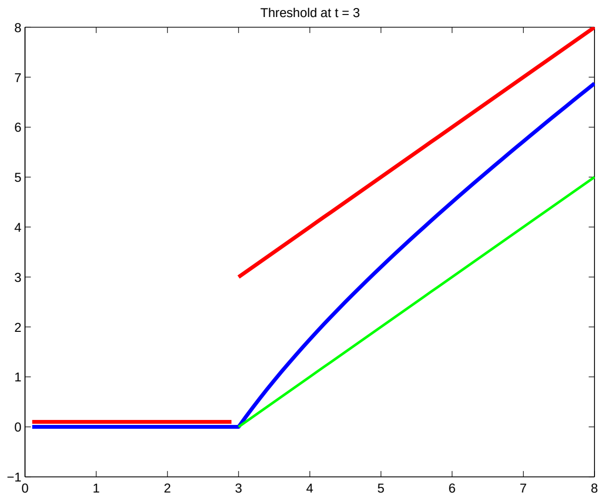
More general **threshold shrinkage** rule:

odd: $\eta(-x, t) = -\eta(x, t)$

shrinks: $\eta(x, t) \le x$ if $x \ge 0$

bounded: $x - \eta(x, t) \le t + b$

threshold: $\eta(x, t) = 0$ iff $|x| \le t$



For example: • $\eta(x, t) = \left(1 - \frac{t^2}{x^2}\right)_+ x$ (Breiman's n.n. garrote)

• posterior *median* (later)

Two (somewhat) uncoupled issues

- choice of *non-linearity*: problem dependent, e.g.
  - hard thresh: – preserves peak heights, good L_2 error
  - soft thresh: – smoother visual appearance
  - intermediate choices

- choice of *threshold value* – Dichotomy:
  - subjective/previous experience (e.g. 3σ - 5σ)
  - vs. *automatic*

Focus here on **threshold choice**

Goals for a Thresholding Method

- data adaptive (to sparse **and** dense sequences)
- stable
- computable, with software
- good performance
  - on simulations
  - on data
  - in theory

E-Bayes on sparse mixtures is **one** solution:
near minimax across several goals:
{theory, simulation, data, software }

This talk: how and why E-Bayes works
theoretical basis for adaptivity and robustness claims

Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Ad Hoc Mixture Model

Prior: $\mu_i \stackrel{i.i.d.}{\sim} f(\mu) = (1-\mathbf{w})\delta_0(\mu) + \mathbf{w}\gamma(\mu)$

i.e.

$$X_i \sim \begin{cases} \phi(x) & \text{w. prob} \quad 1-\mathbf{w} \\ g(x) = \int \phi(x-\mu)\gamma(\mu)d\mu & \text{w. prob} \quad \mathbf{w} \end{cases}$$

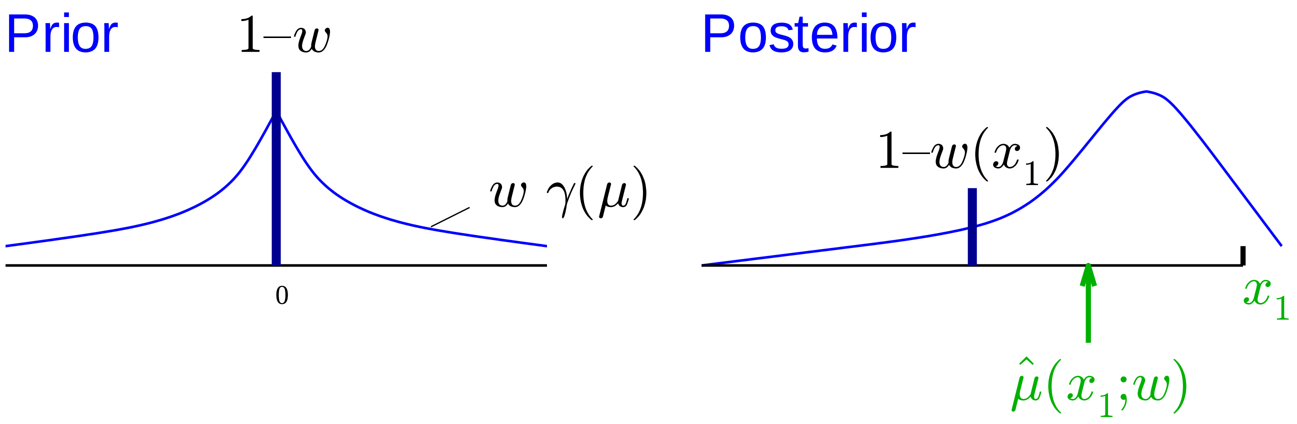
Given $\mathbf{w}$, posteriors for μ_i are **independent**

$\Rightarrow$ Bayes estimates for ℓ_q loss:

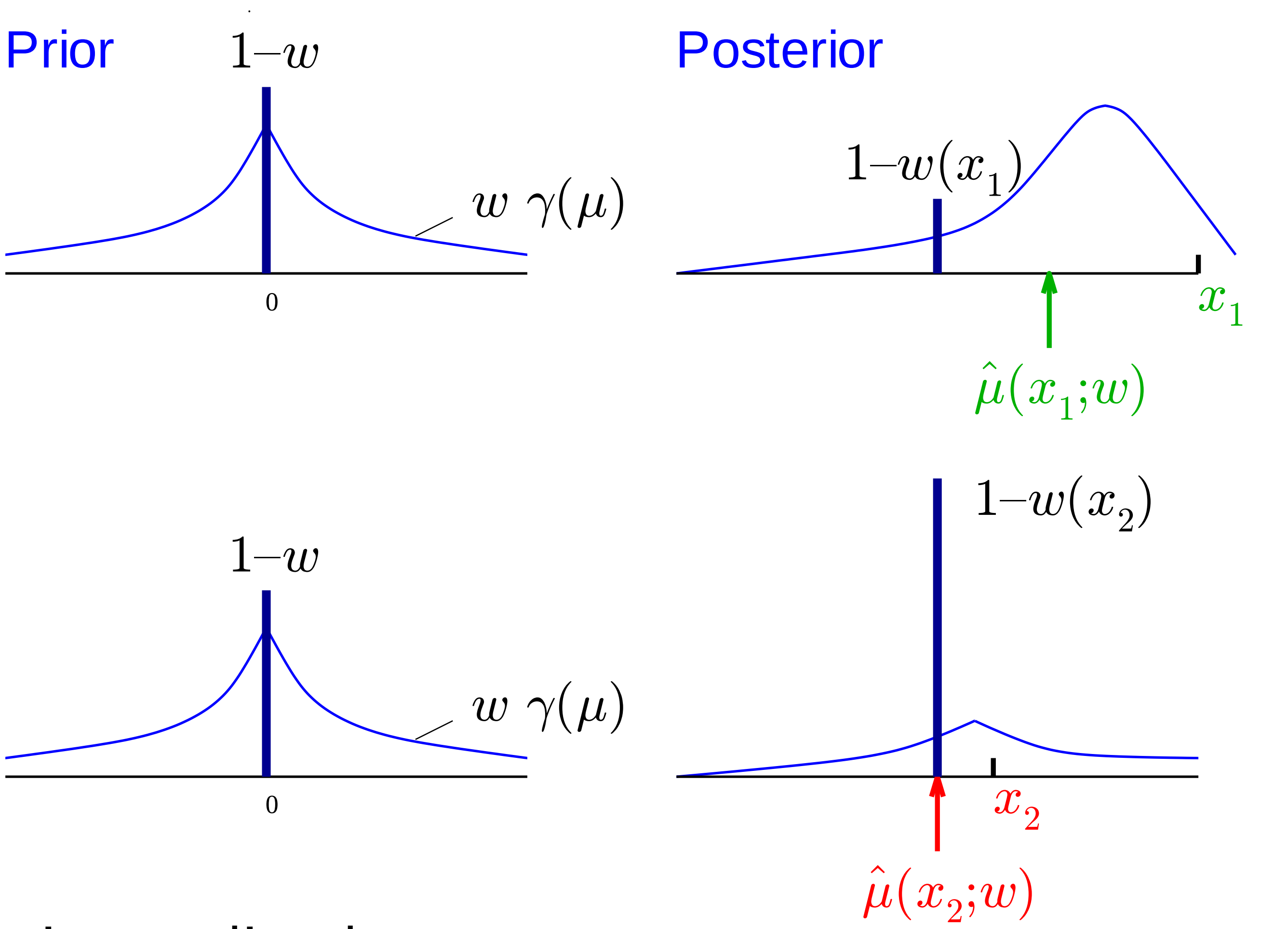
$$\hat{\mu}_i(x) = \hat{\mu}(x_i; w) = \text{argmin}_a E[|\mu_i - a|^q | x_i]$$

Posterior mean/**median**/mode for $q = 2/\mathbf{q = 1}/q = 0$

Posterior median does thresholding



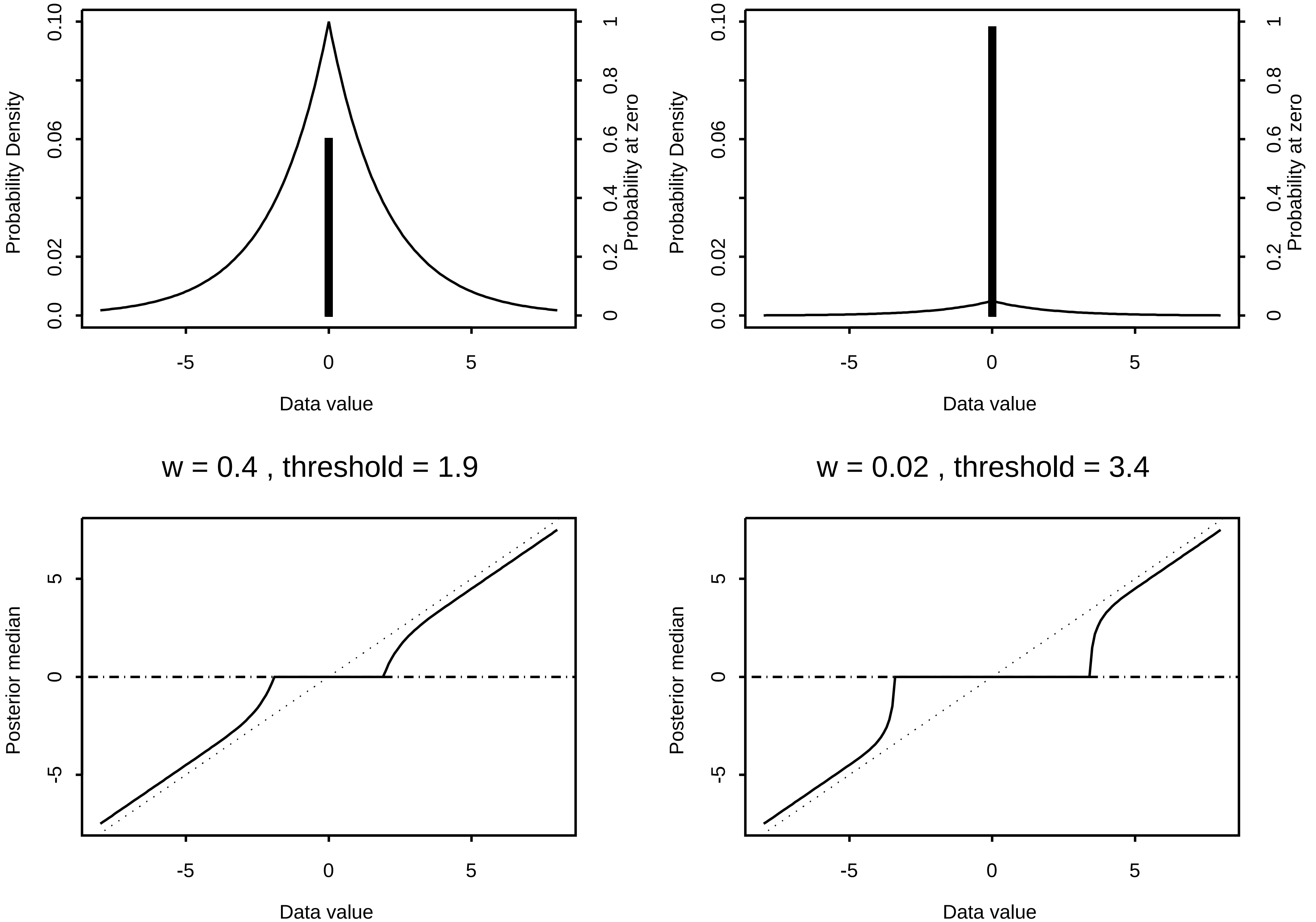
Posterior median does thresholding



$\Rightarrow$ posterior median has:

- threshold zone: $\hat{\mu}(x;w) = 0 \Leftrightarrow |x| \leq t(w)$
- shrinkage: $|\hat{\mu}(x;w)| \leq x$

Two Specific Examples



Key properties

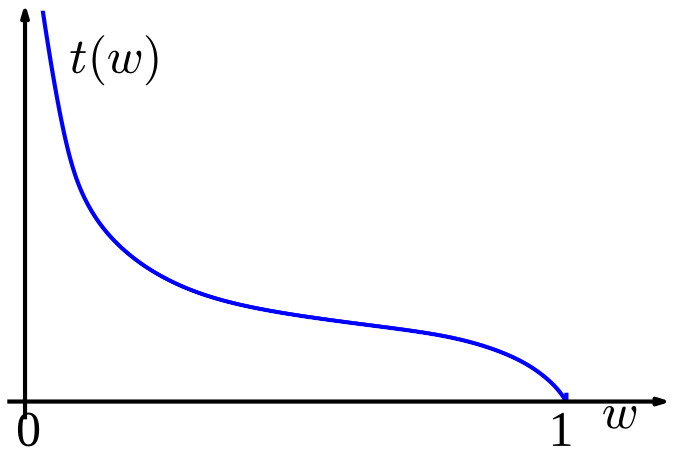
Typical priors: $\gamma(\mu) = \frac{1}{2} a e^{-a|\mu|}$ [Laplace]

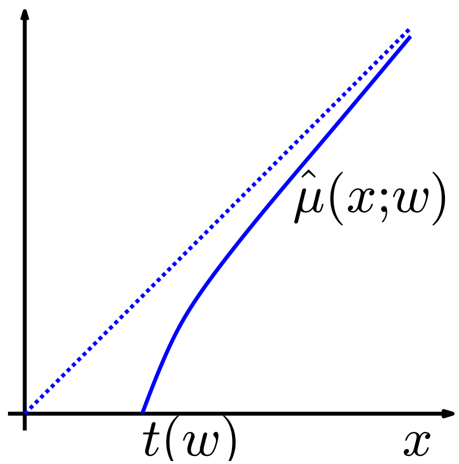
$$\sim \begin{cases} N(0, \theta^{-1} - 1) \\ \theta \sim Beta(\frac{1}{2}, 1) \end{cases}$$ [Quasi-Cauchy]

Consequences:

- Small $w \Leftrightarrow$ Large threshold $t(w)$
- Bounded shrinkage property: $\exists b$ s.t. for all x, w

$$|x - \hat{\mu}(x; w)| \le t(w) + b$$





Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Estimation of w: Marginal Max. Likelihood

Log-likelihood: $\ell(w) = \sum_{i=1}^{n} \log\{(1-w)\phi(X_i) + wg(X_i)\}$

Score function:

$$S(w) = \ell'(w) = \sum_i \frac{g(X_i) - \phi(X_i)}{\phi(X_i) + w[g(X_i) - \phi(X_i)]}$$

$$= \sum_i \frac{\beta(X_i)}{1 + w\beta(X_i)},$$

where the mixture ratio

$$\beta(x) = \frac{g(x)}{\phi(x)} - 1. \qquad\qquad (g = \gamma \star \phi)$$

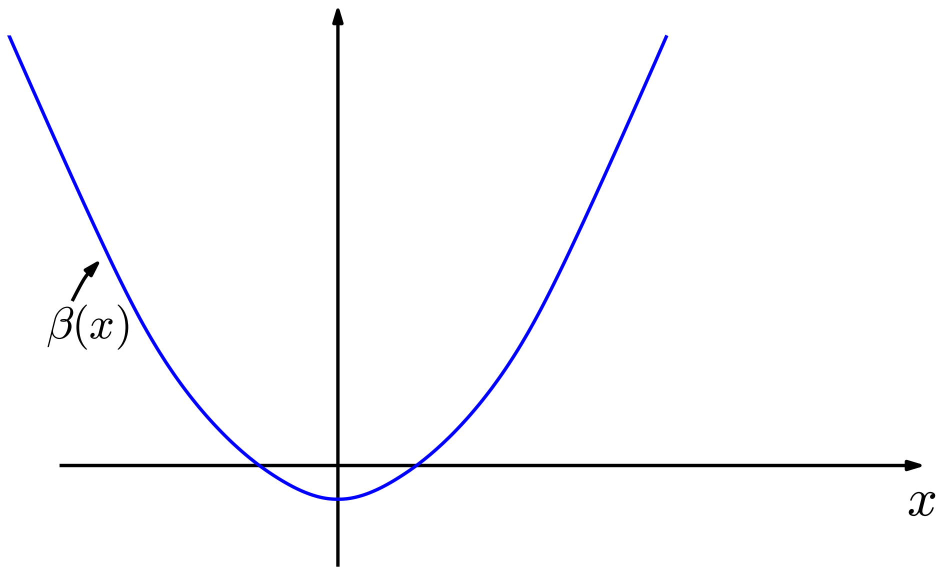
$S(w)$ is monotone $\searrow$, so estimate w from $S(\hat{w}) = 0$.

$\rightarrow$ **"E-Bayes" estimate**: $\qquad \hat{t} = t(\hat{w})$.

Why MML **might** fail: Heavy tails of $\beta(x)$

$$\beta(x) = \frac{g(x)}{\phi(x)} - 1; \qquad S(w) = \sum_i \frac{\beta(X_i)}{1 + w\beta(X_i)}$$

E.g. $\gamma(\mu) = \frac{a}{2} e^{-a|\mu|}$, "Laplace" $\Rightarrow$ $\beta(x) \asymp e^{(x-a)^2/2}$

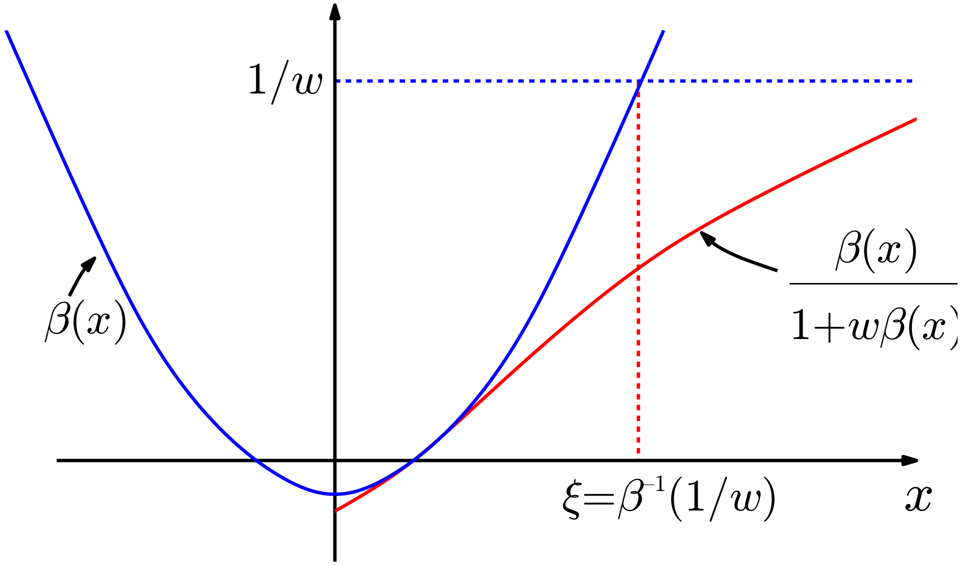


Infinite Variance! Var $S(0) = \infty$, since Var $\beta(X) = \infty$

Why MML **might** fail: Heavy tails of $\beta(x)$

$$\beta(x) = \frac{\gamma \star \phi(x)}{\phi(x)} - 1; \qquad S(w) = \sum_i \frac{\beta(X_i)}{1 + w\beta(X_i)}$$

E.g. $\gamma(\mu) = \frac{a}{2} e^{-a|\mu|}$ "Laplace" $\Rightarrow$ $\beta(x) \asymp e^{(x-a)^2/2}$



Infinite Variance! Var $S(0) = \infty$, since Var $\beta(X) = \infty$

but

w regularises! Var $S(w) \asymp n \cdot \frac{1}{w}$ $\nearrow \infty$ as $w \searrow 0$.

Why MML **works**: Heavy tails of $\beta(x)$

$$S(\hat{w}) = \sum_i \beta(x_i, \hat{w}) = 0$$

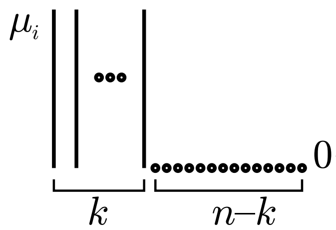
Note: at $\hat{w}$, bimodality of $\beta(x_i, \hat{w})$:

$$\beta(x, \hat{w}) = \frac{\beta(x)}{1 + w\beta(x)}$$

$$\approx \begin{cases} 1/w & \beta(x) \text{ large} \\ \beta(x) & \beta(x) \text{ small} \end{cases}$$

For simulation examples:

Configuration

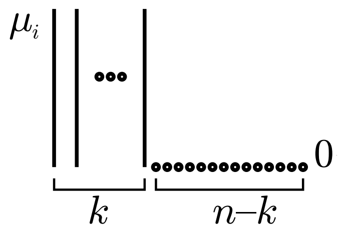


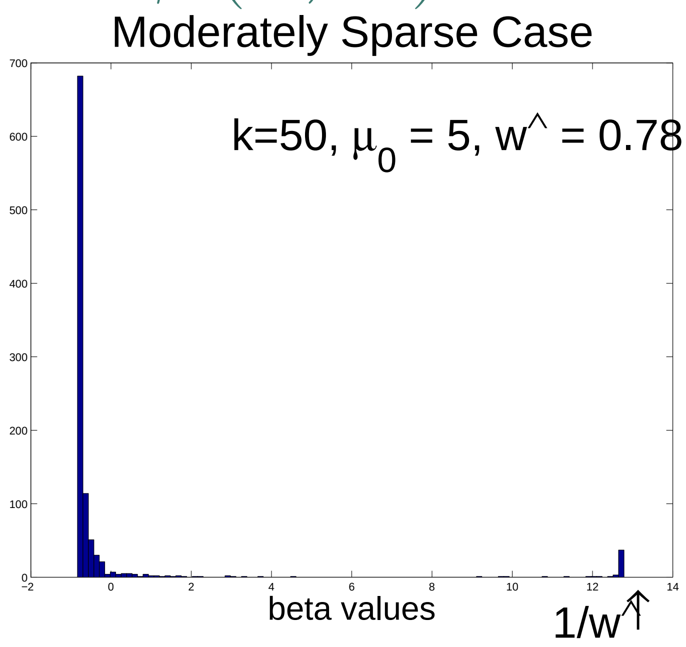
Histograms of $\beta(x, \hat{w})$

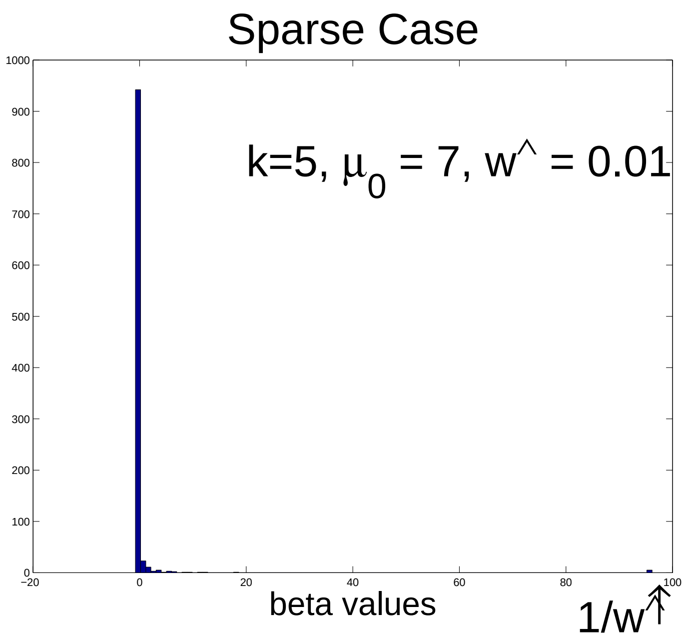
$$\beta(x,\hat{w}) = \frac{\beta(x)}{1 + w\beta(x)}$$

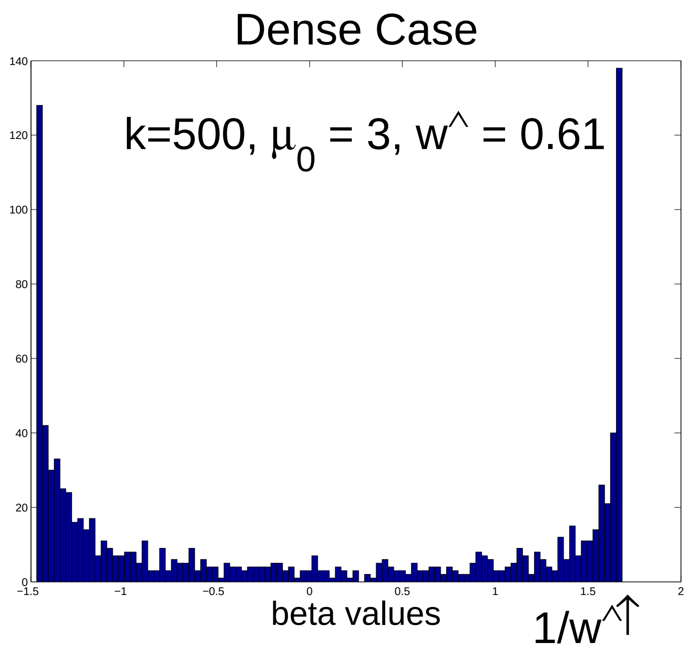
$$\approx \begin{cases} 1/w & \beta(x) \text{ large} \\ \beta(x) & \beta(x) \text{ small} \end{cases}$$

Configuration









Why MML **works**: Heavy tails of $\beta(x)$

Hence

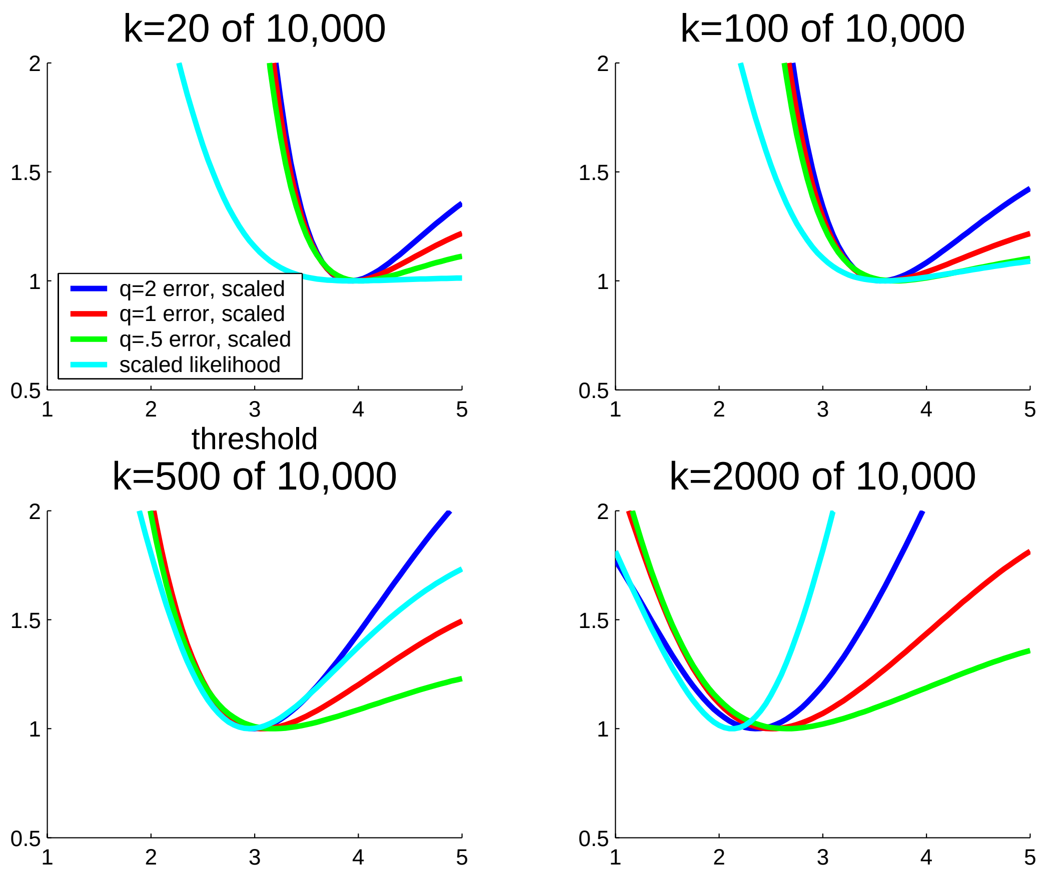
- Unique MMLE $S(\hat{w}) = 0$,
- easy to compute,
- extreme bimodality of $\beta(x_i, \hat{w}) = \frac{\beta(x_i)}{1+\hat{w}\beta(x_i)}$

So heavy tails of β (and monotonicity)

$\Rightarrow$ stability & low variance of $\hat{w}$.

Likelihood vs Risk Minima

- Compare locations of likelihood and loss minima
- q-minima are often close; E-Bayes quite close
- Percent increase from minimum at E-Bayes solution often small



Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Stein's Unbiased Risk Estimate (SURE)

- Choose t to minimize – for **soft** thresholding –

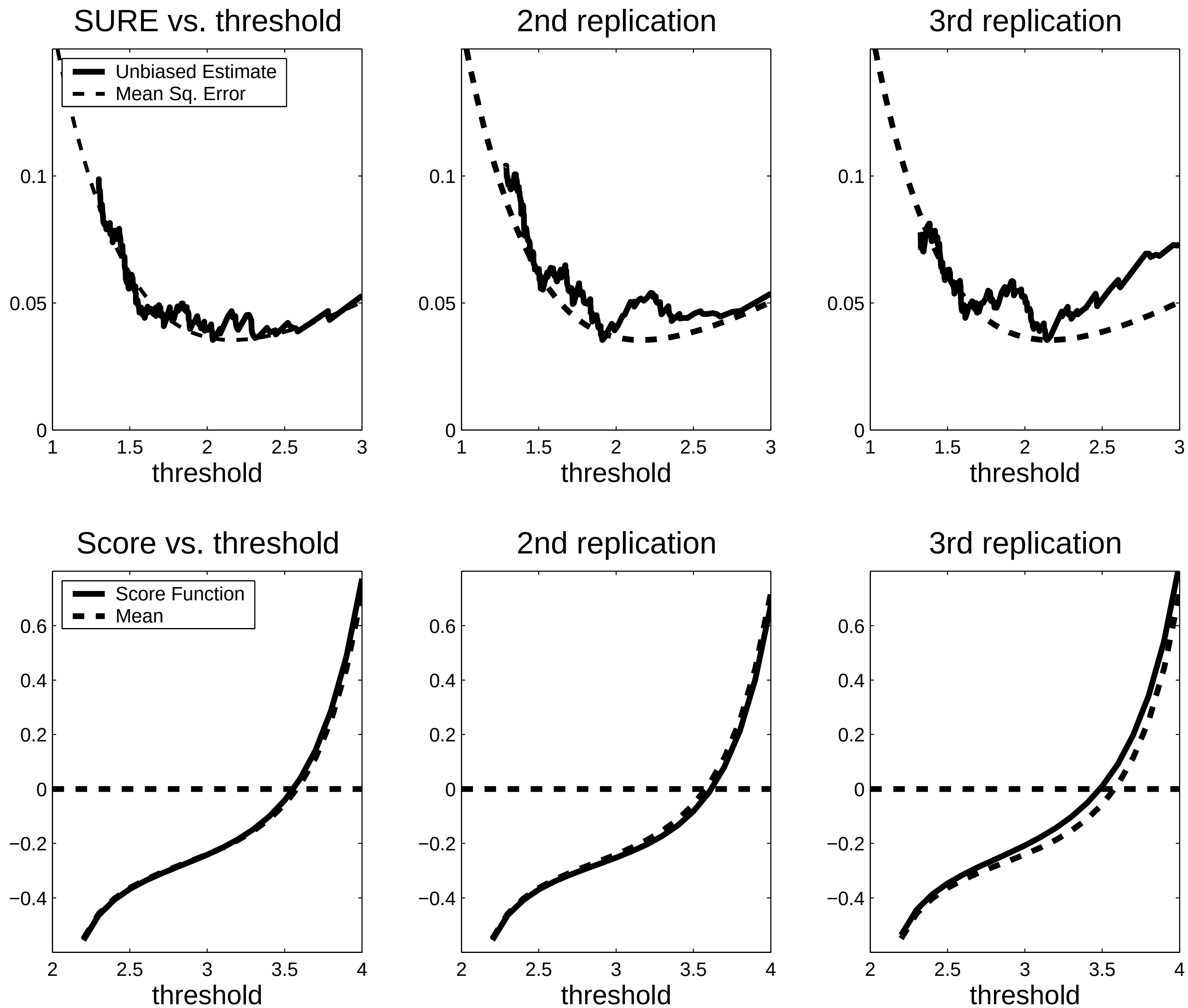
$$\hat{U}(t) = n + \sum_{1}^{n} x_k^2 \wedge t^2 - 2 \sum_{1}^{n} I\{x_k^2 \le t^2\}$$

$$\hat{t}_{SURE} = \underset{0 \le t \le \sqrt{2 \log n}}{\text{argmin}} \; \hat{U}(t)$$

- some good theory, **but** unstable **and**
- doesn't handle sparse cases well

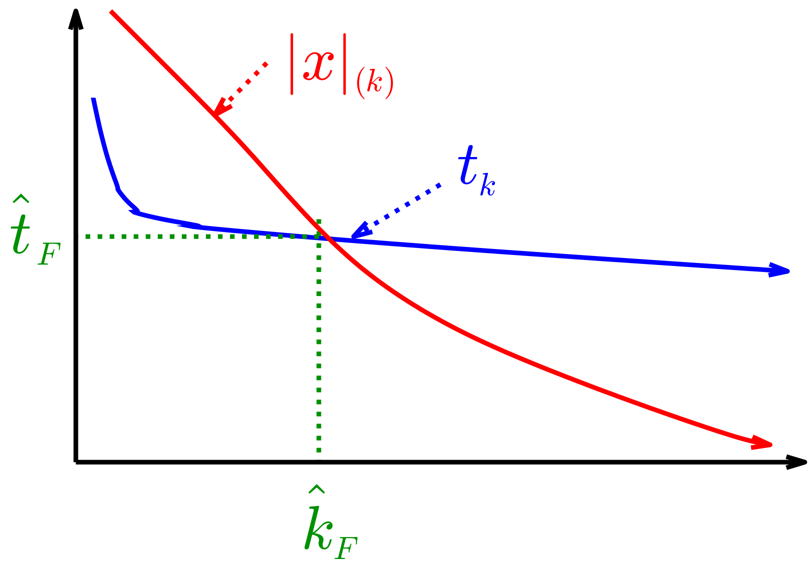
- instability is inherent to SURE on thresholds
  - similar plots for SURE for *posterior medians*

Unbiased risk vs Score criteria



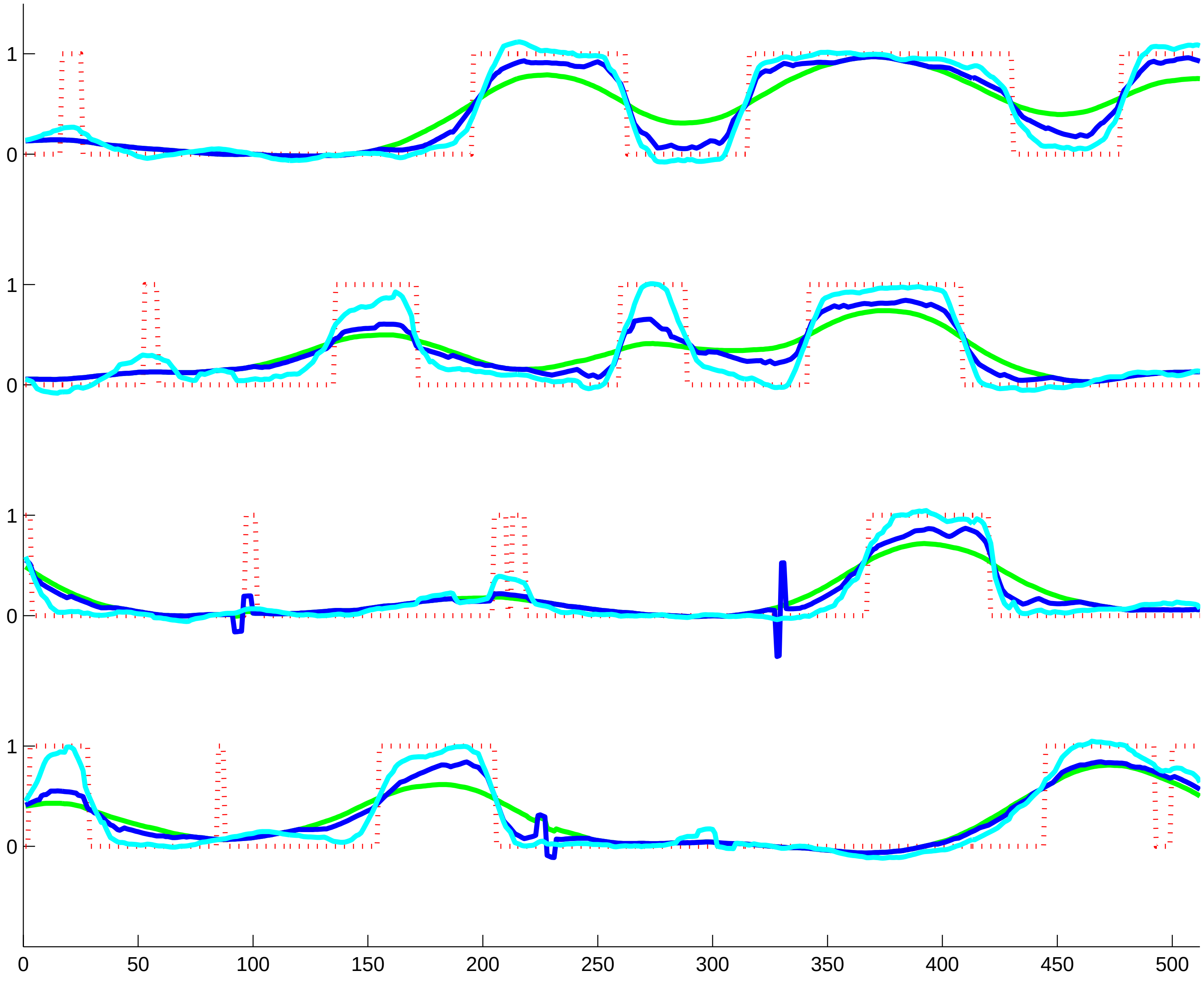
False Discovery Rate (FDR)

- decreasing magnitudes $|x|_{(1)} \geq |x|_{(2)} \geq \ldots \geq |x|_{(n)}$
- quantile boundary $t_k = \sigma z(\frac{q}{2} \cdot \frac{k}{n})$
- FDR parameter $q \in (0, 1/2]$
- crossing index $\hat{k}_F = \max\{k : |x|_k \geq t_k\}$

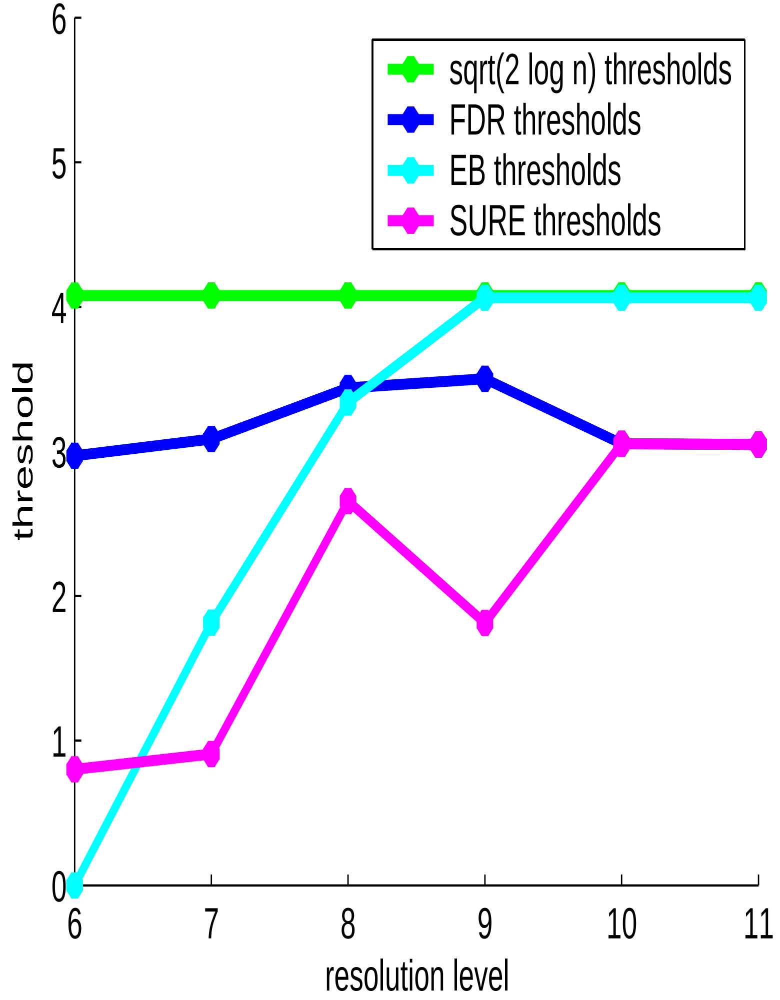


N.B.: doesn't handle **dense** signals well (for q small):

E-Bayes thresholds: 2.9% errors



FDR chooses large t in dense cases



- $\sqrt{2\log n}$ too big at coarse j
- SURE too low at fine j
- FDR(.05) still too big at coarse j
- EBayes: good transition across scales

Where we're headed (i)

Claim: Sparse-mixture E-Bayes thresholding is near-optimal over several goals

$\{$ **theory**, simulation, data, software $\}$

Last three: see IMJ & B.W. Silverman, AOS, 2004 in press,

- "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences",
- "Empirical Bayes selection of wavelet thresholds",
- and **numerous references to other work** therein!

and available software, at

- `www.stats.ox.ac.uk/~silverma/`

Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Where we're headed (ii)

- Presence of **phase changes**

  $$\textit{v.sparse} \quad \longrightarrow \quad \textit{sparse} \quad \longrightarrow \quad \textit{dense}$$

  challenges good threshold selection in sequence model.
- not all *a priori* reasonable methods are in fact satisfactory in adapting over the range of phases
- Some methods (**E-Bayes**, penalized least squares ...) can be shown to do so
- (adaptation to) phase changes in function estimation

  $$\textit{Regular} \quad \longrightarrow \quad \textit{Critical} \quad \longrightarrow \quad \textit{Logarithmic}$$

  flows from phase changes in (growing) Gaussian sequence models

Theory for adaptive thresholds

Goals:

- **flexible adaptation** small risk for small $\|\mu\|_p$
- **robustness** bounded risk for all μ
- **variety of losses** $\ell_q, 0 < q < 2$

A priori – not clear for MML:

- **know** that prior is wrong: μ is arbitary
- how does $\mathcal{L}(t(\hat{w}))$ depend on μ?

Phase changes in minimax risk

Region

Non-zero signal

Typical near-mmx estimator

Minimax risk for ℓ_p balls

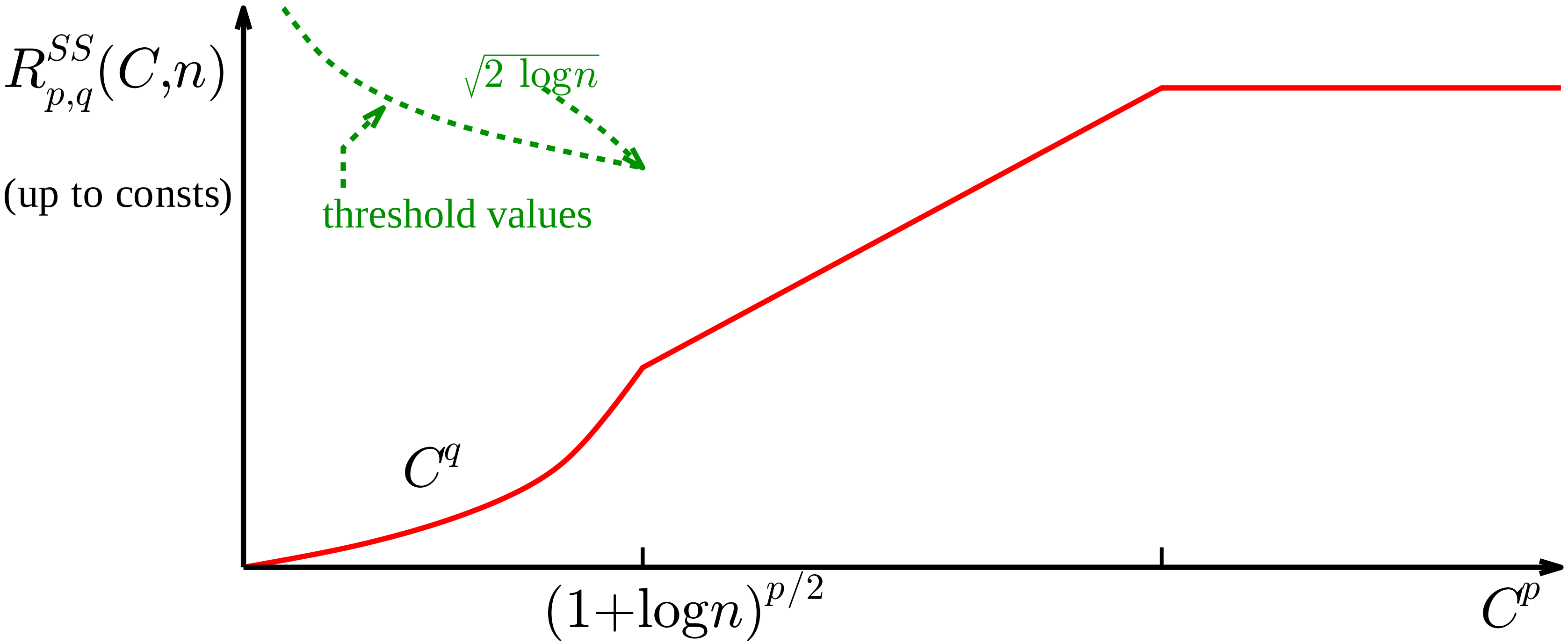
- a family of benchmarks, using
  - $\ell_p(C) = \{\mu : \sum_1^n |\mu_i|^p \le C^p\}$
  - p - enforces sparsity (when < 2)
  - C - measures size (and, indirectly, sparsity)
- using ℓ_q loss $(0 < q \le 2)$

$$R_{p,q}^{SS}(C,n) := \inf_{\hat{\mu}} \sup_{\mu \in \ell_p(C)} E \sum_1^n |\hat{\mu}_i - \mu_i|^q$$

- increases from 0 to $nE|Z|^q$ with C from 0 to ∞
- want $\hat{\mu}$ to adapt to unknown p, C (for all q)
- an essential feature: **'phase changes'**:

v.sparse $\longrightarrow$ *sparse* $\longrightarrow$ *dense*

Phase changes in minimax risk



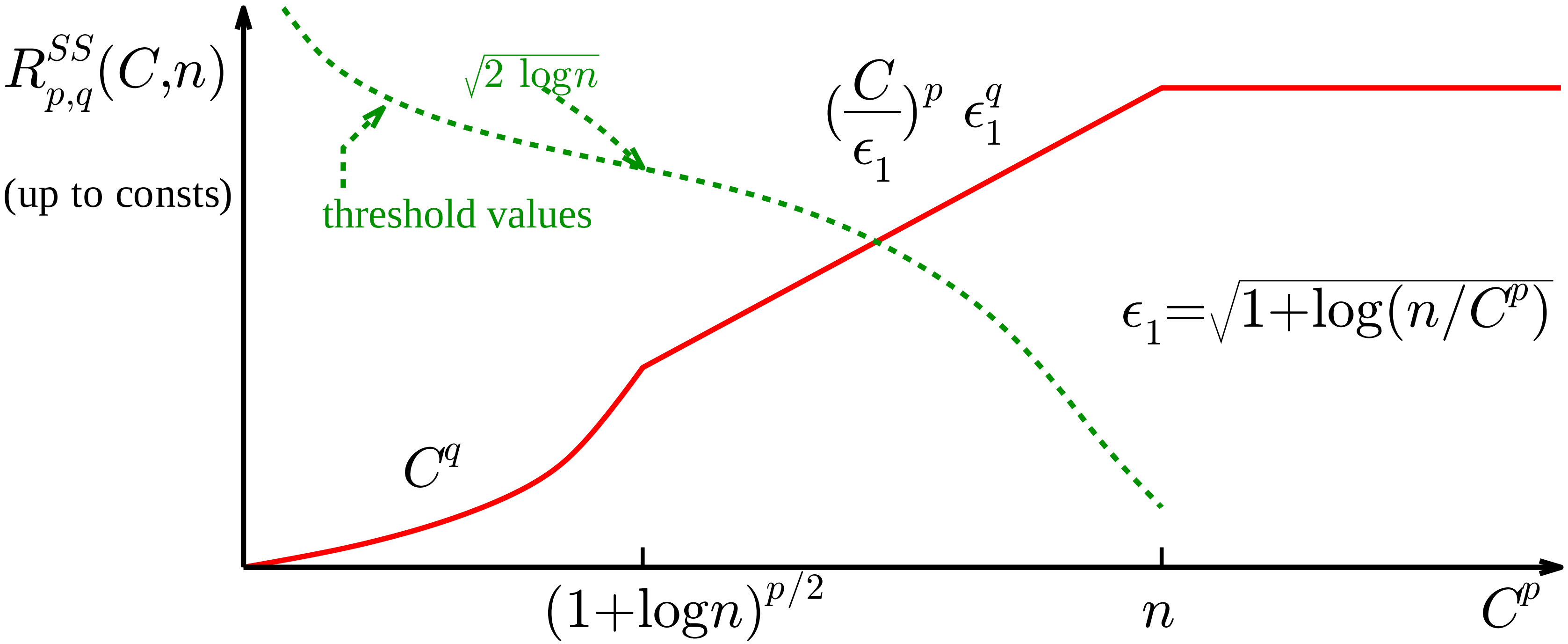
Region v. sparse

Non-zero signal $\big|^{C}$

$\sharp = 1$

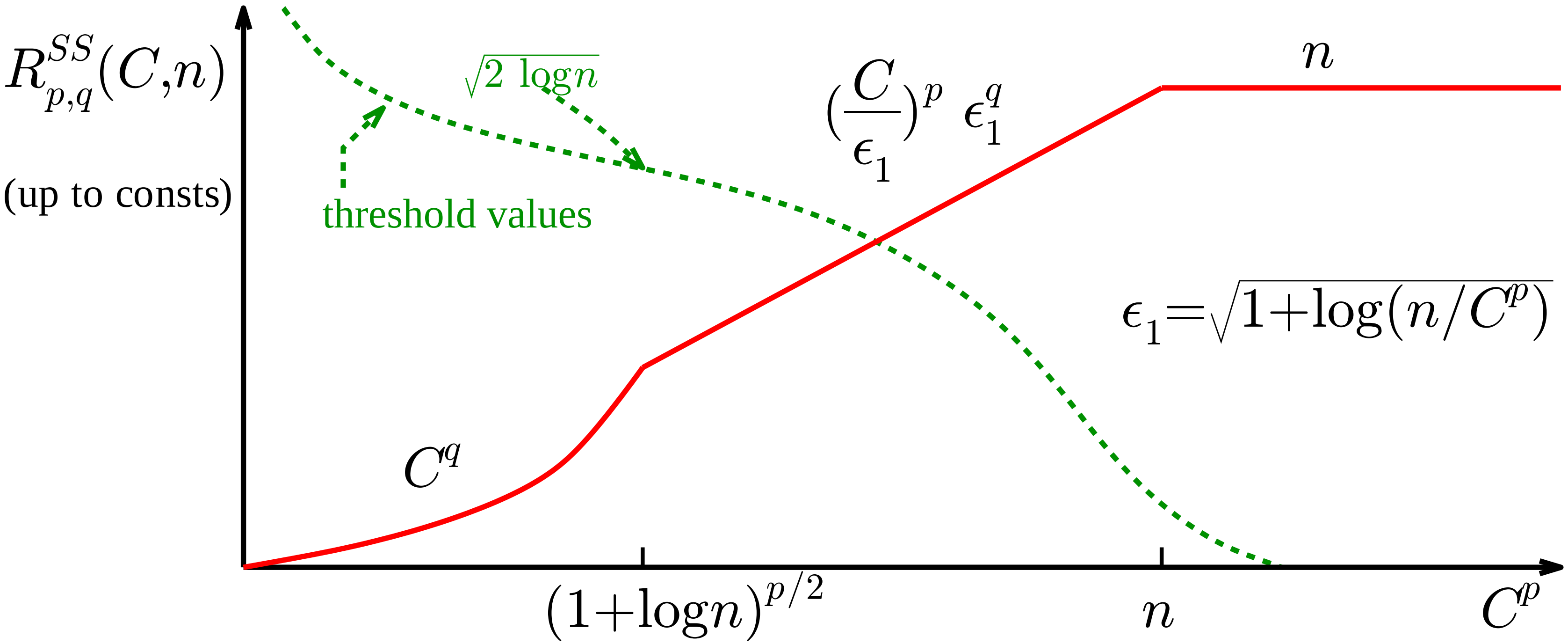
Typical near-mmx estimator $\hat{\theta}=0$

Phase changes in minimax risk



Region	v. sparse	sparse
Non-zero signal	C	$\cdots$ ϵ_1
	$\sharp = 1$	$\sharp \asymp (C/\epsilon_1)^p$
Typical near-mmx estimator	$\hat{\theta} = 0$	threshold at ϵ_1

Phase changes in minimax risk



Region	v. sparse	sparse	dense
Non-zero signal	C	$\cdots$ ϵ_1	$\cdots\cdots\cdots\cdots$ ϵ_1
	$\sharp = 1$	$\sharp \asymp (C/\epsilon_1)^p$	$\sharp \asymp n$
Typical near-mmx estimator	$\hat\theta = 0$	threshold at ϵ_1	$\hat\theta_{MLE}(y) = y$

Bounds for estimated threshold $\hat{t}$:

Response to sparsity: $\|\mu\|_p$ **small** $\Rightarrow \hat{t}$ **'large'**

With high probability, uniformly over $n^{-1} \sum |\mu_i|^p \leq \eta^p$,

$$\hat{t}^2 > \begin{cases} 2 \log \eta^{-p} + (p-1) \log \log \eta^{-p} & \text{if } \eta^p \geq \frac{\log^2 n}{n} \\ 2 \log n \quad - (5-p) \log \log n & \text{if } \eta^p < \frac{\log^2 n}{n} \end{cases}$$

Response to 'density': $\|\mu\|_0$ **large** $\Rightarrow \hat{t}$ **'small'**

Density exceeds π on $D_\tau(\pi) = \{\mu : n^{-1} \#\{i : |\mu_i| \geq \tau\} \geq \pi\}$

With high probability, uniformly over $\mu \in D_\tau(\pi)$

$$\hat{t} \leq t(\tau, \pi)$$

where $t(\tau, \pi) \searrow$ as $\pi \nearrow$.

[Actually true for "pseudo-threshold" $\xi = \beta^{-1}(1/w)$]

Main Adaptivity result for E-Bayes(A)

For $n \geq n_0$ and all $0 < p, q \leq 2$ and all $C > 0$,

$$\sup_{\mu \in \ell_p(C)} E\|\hat{\mu}_{EB} - \mu\|_q^q \leq c\big[R_{p,q}^{SS}(C,n) + n^{-A}(\log n)^{(q-1)/2}\big]$$

[if $A = 0$, same result with error $(\log n)^{2+(q-p)/2}$]

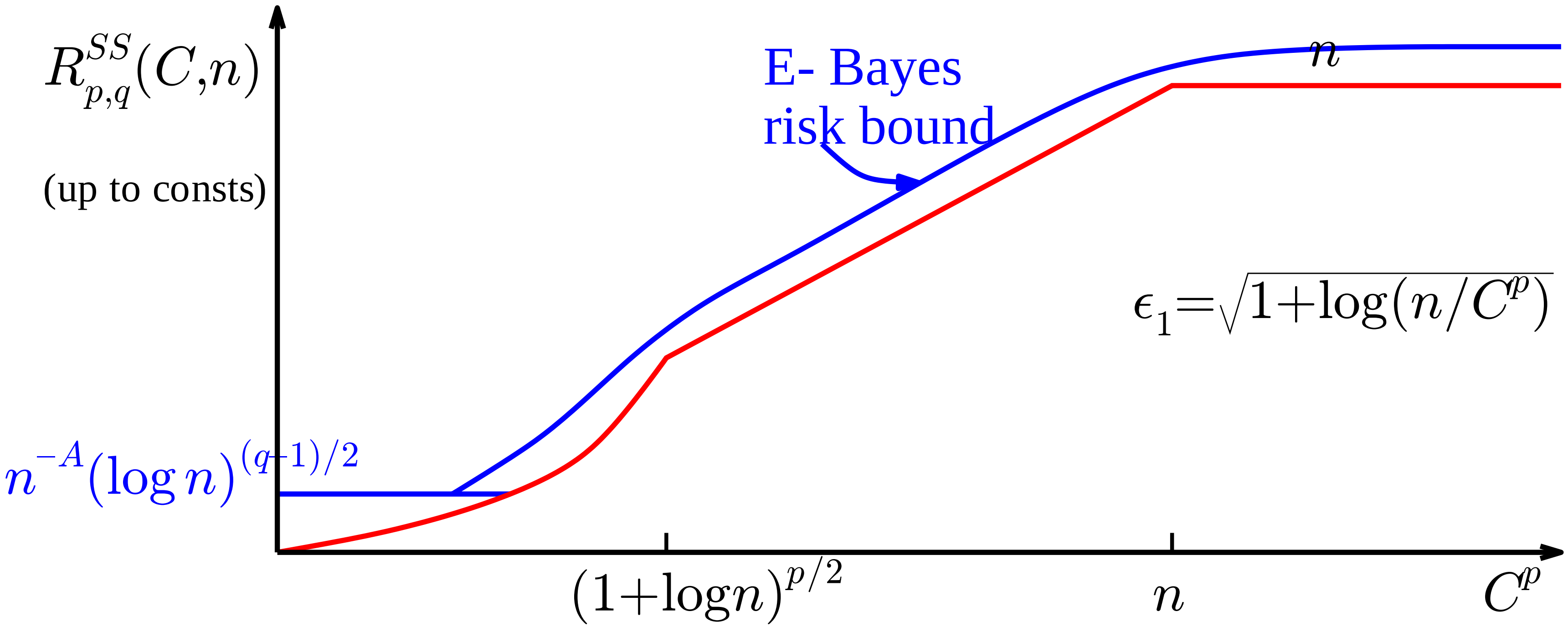
Applies to **any** estimator $\hat{\mu}_i(x, \hat{w}) = \eta(x_i, t(\hat{w}))$ if

- $\eta(x, t)$ is bounded threshold shrinkage rule
- $\hat{w}$ is chosen by Empirical Bayes

E-Bayes(A) To ensure v. small mean ℓ_q error when μ small, need thresholds $> \sqrt{2 \log n}$. Theory suggests an **ad hoc** fix: let $A > 0$, and

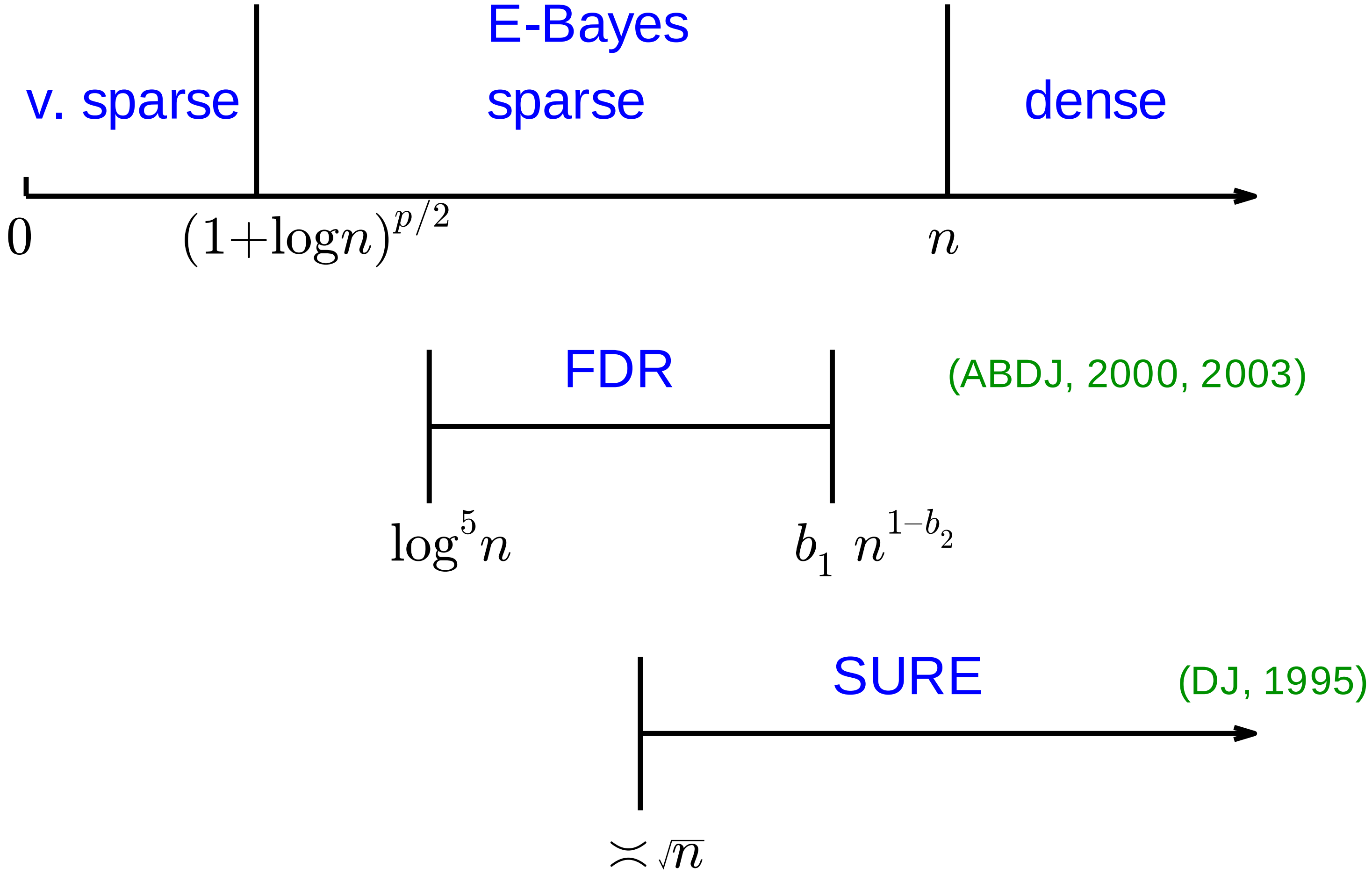
$$\hat{t}_A = \begin{cases} \hat{t} & \text{if } \hat{t}^2 \leq 2\log n - 5 \log\log n \\ \sqrt{2(1+A)\log n} & o/w \end{cases}$$

E-Bayes Adaptation to Phase Changes



Region	v. sparse	sparse	dense
Non-zero signal	C	$\cdots$ ϵ_1	$\cdots\cdots\cdots\cdots$ ϵ_1
	$\sharp = 1$	$\sharp \asymp (C/\epsilon_1)^p$	$\sharp \asymp n$
Typical near-mmx estimator	$\hat\theta = 0$	threshold at ϵ_1	$\hat\theta_{MLE}(y) = y$

Contrast with some existing results



[Birgé - Massart (2001): adaptivity for ℓ_2 loss; non-asymptotic bounds; *via* complexity penalized least squares; C-H Zhang: non-parametric E-Bayes]

Agenda

- Transform shrinkage & block structure
  - why data-dependent thresholds matter
- Reduction to Single Sequence (Growing Gaussian) model
- An *ad hoc* mixture model
  - posterior median thresholding
  - Empirical Bayes threshold choice
  - comparison with other methods (SURE, FDR)
- Adapting to phase changes in the GG model
- Adapting to phase changes in wavelet shrinkage

Function Estimation

Back to

$$y_i = f(i/n) + \sigma\epsilon_i \qquad \text{(non-parametric reg'n)}$$

or

$$Y_t = \int_0^t f(s)ds + N^{-1/2}W_t \qquad \text{(Gaussian White Noise)}$$

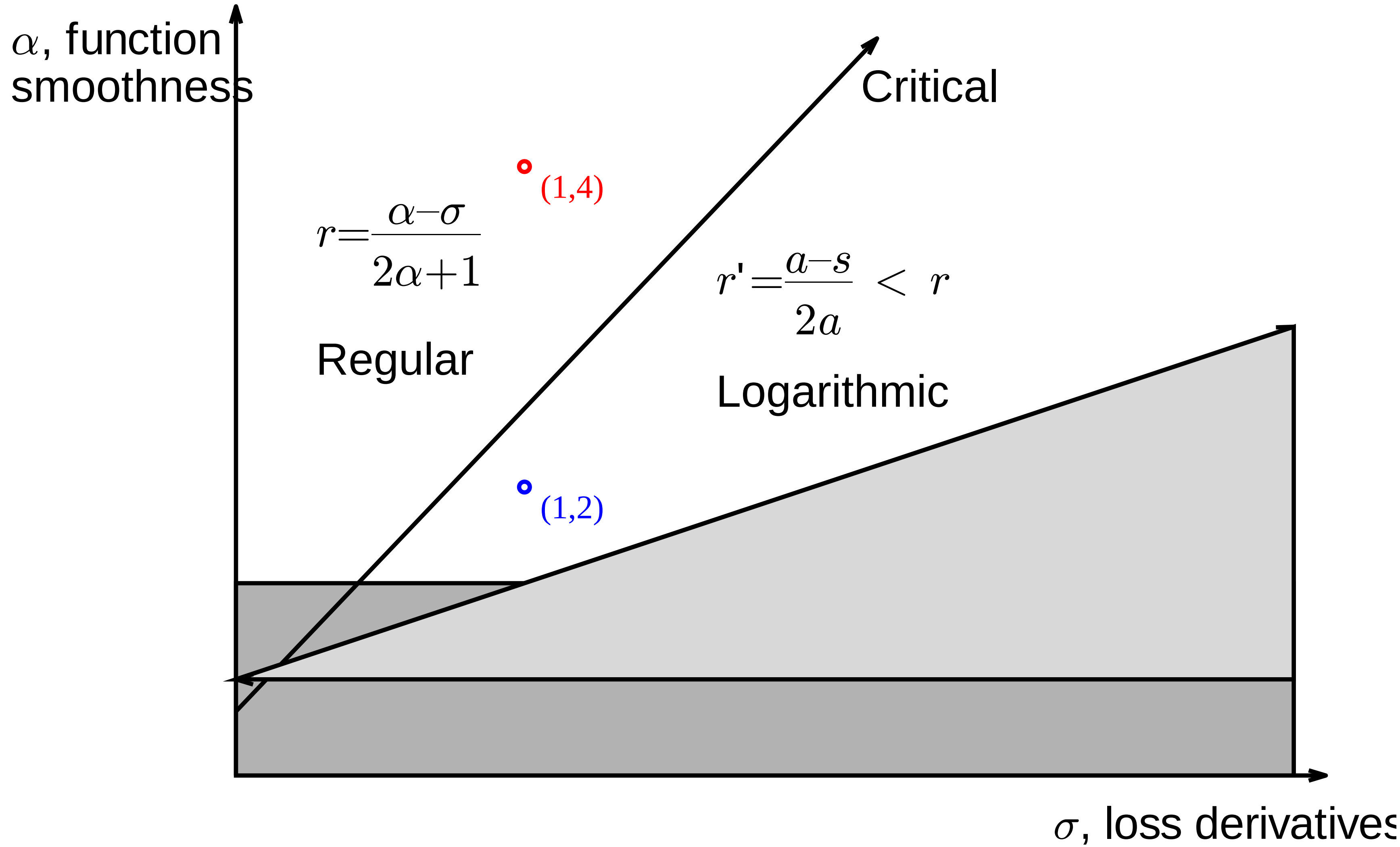
or

$$y_{jk} = \theta_{jk} + N^{-1/2}z_{jk} \qquad \text{(wavelet coefficients)}$$

Goal: apply results for single sequence model to each (wavelet) level

Issue: consequences of phase changes?

Phase Change in Rates: $(q > p)$



Assumptions

Smoothness: $\alpha \leftrightarrow \#$ derivatives for f $\qquad (\alpha > 1/2 - 1/p)$
$p \leftrightarrow$ index of average smoothness, $p \leq \infty$

$$\theta \in B^{\alpha}_{p,\infty}(C) : \qquad \sum_{k=1}^{2^j} |\theta_{jk}|^p \leq C^p 2^{-ap}, \qquad \forall j, \quad a = \alpha + 1/2 - 1/p$$

Error Measures: $\sigma \leftrightarrow \#$ derivatives estimated, $0 \leq \sigma < \alpha - (1/p - 1/q)$
$q \leftrightarrow$ index of average error, $\quad q \leq 2$

$$E\|\hat{\theta} - \theta\|^q_{B^{\sigma}_{q,q}} = \sum_j 2^{sqj} E\|\hat{\theta}_j - \theta_j\|^q_q, \qquad s = \sigma + 1/2 - 1/q$$

$\sim$ estimation of θ^{th} derivative in L_q:

$$a_0 \|\theta\|^q_{B^{\sigma}_{q,2}} \leq \int |f^{(\sigma)}|^q \leq a_1 \|\theta\|^q_{B^{\sigma}_{q,q}}$$

Phase Change in Rates

Assume $\qquad a > 0, \sigma \geq 0, \alpha > \sigma + (1/p - 1/q)_+$

Rates:

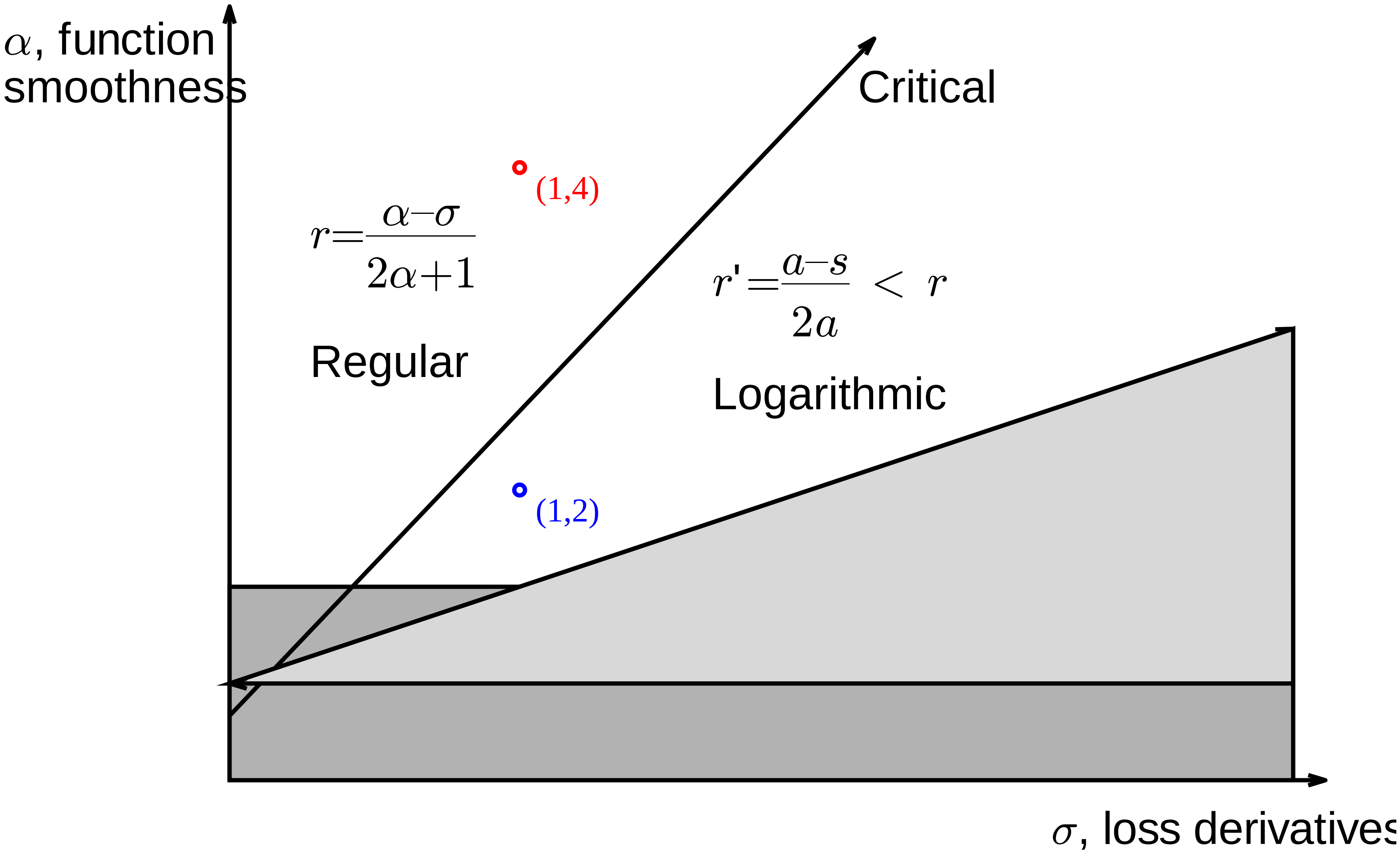
$$\mathbf{r} = \frac{\alpha - \sigma}{2\alpha + 1}, \qquad \mathbf{r'} = \frac{a - s}{2a}$$

Minimax risk (DJKP, 95):

$$R(\Theta(C), N) = \inf_{\hat{\theta}} \sup_{B^{\alpha}_{p,\infty}(C)} E\|\hat{\theta} - \theta\|^q_{B^{\alpha}_{q,q}}$$

$$\asymp \begin{cases} C^{(1-2r)q}\left(\frac{1}{N}\right)^{\mathbf{r}q} & \text{on } \textbf{Regular} \quad (\mathcal{R}) \\ C^{(1-2r')q}\left(\frac{\log N}{N}\right)^{\mathbf{r'}q} & \text{on } \textbf{Logarithmic} \quad (\mathcal{L}) \\ C^{(1-2r')q}\frac{(\log N)^{\mathbf{r'}q+1}}{N^{\mathbf{r'}q}} & \text{on } \textbf{Critical} \quad (\mathcal{C}) \end{cases}$$

Phase Change in Rates: $(q > p)$



$$[a = \alpha + 1/2 - 1/p; \qquad s = \sigma + 1/2 - 1/q]$$

Levelwise (Empirical Bayes) Thresholding

For $y_j =$

empirical wavelet coefficients of data (at j^{th} level)

or

data in dyadic form of Gaussian white noise model
$y_{jk} = \theta_{jk} + N^{-1/2} z_{jk} \qquad j \geq 0, k = 1, \ldots, 2^j.$

use

variance-1 single sequence E-Bayes(A); with
w estimated separately at each level j:

$$\hat{\theta}_j^{EB} = N^{-1/2} \hat{\mu}_{EB}(N^{1/2} y_j; \hat{w}_j)$$

[If $j \geq \log_2 N$, set $\hat{\theta}_j^{EB} = 0$.]

Main Adaptivity Result for E-Bayes(A)

If $0 < p, q \leq 2, sq \leq A$, then for any $\Theta(C)$ in $\mathcal{R} \cup \mathcal{L} \cup \mathcal{C}$

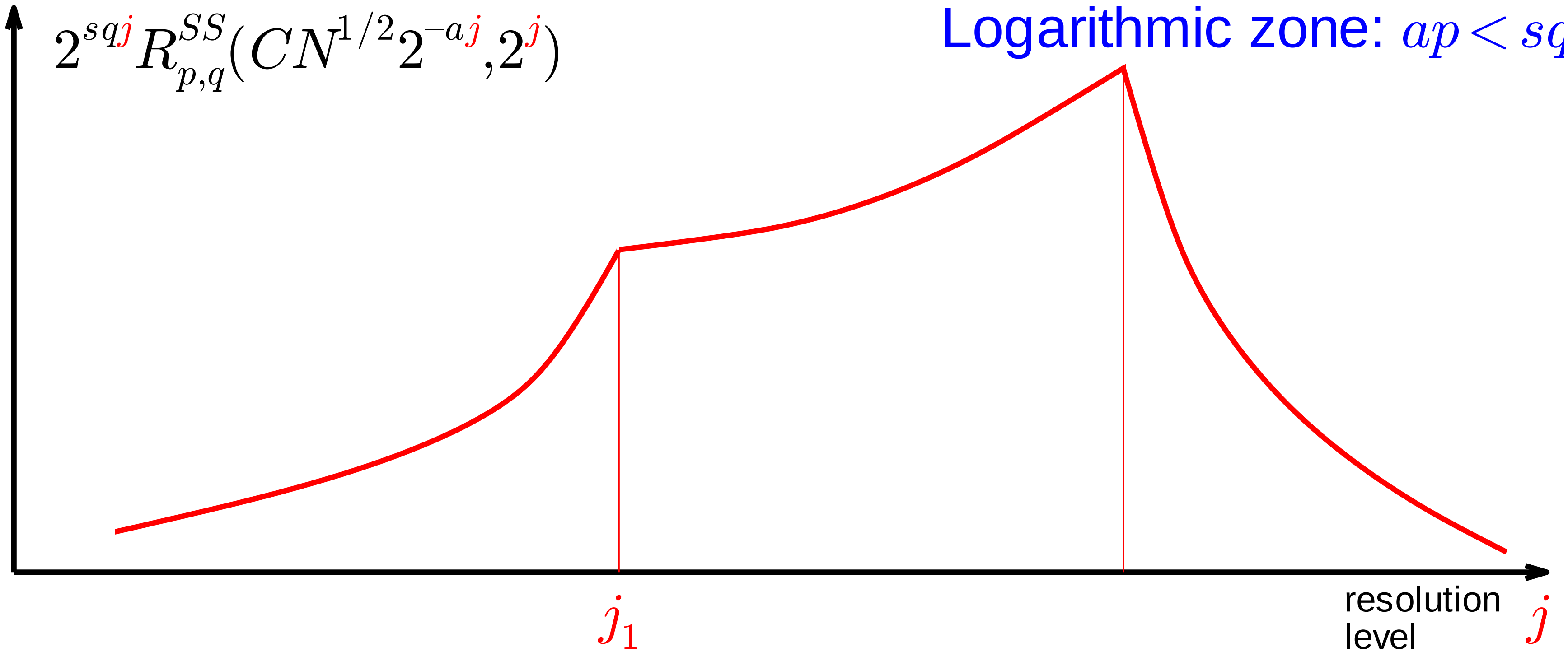
$$\sup_{\theta \in \Theta(C)} E\|\hat{\theta}^{EB} - \theta\|^q_{B^\sigma_{qq}}$$

$$\leq c\big[\underbrace{R(\Theta(C), N)}_{\text{minimax risk}} + \underbrace{C^q N^{-r''q} + N^{-q/2} \log^\nu N}_{\text{lower order}}\big]$$

$r'' = \alpha - \sigma - (1/p - 1/q)_+ \geq \min(r, r'); \ 0 \leq \nu \leq 4.$

i.e. E-Bayes(A) attains optimal rate for full range of **smoothness** α and **losses** σ.

Single Sequence Phase Diagram transforms to:



Reduction to Single Sequence Model

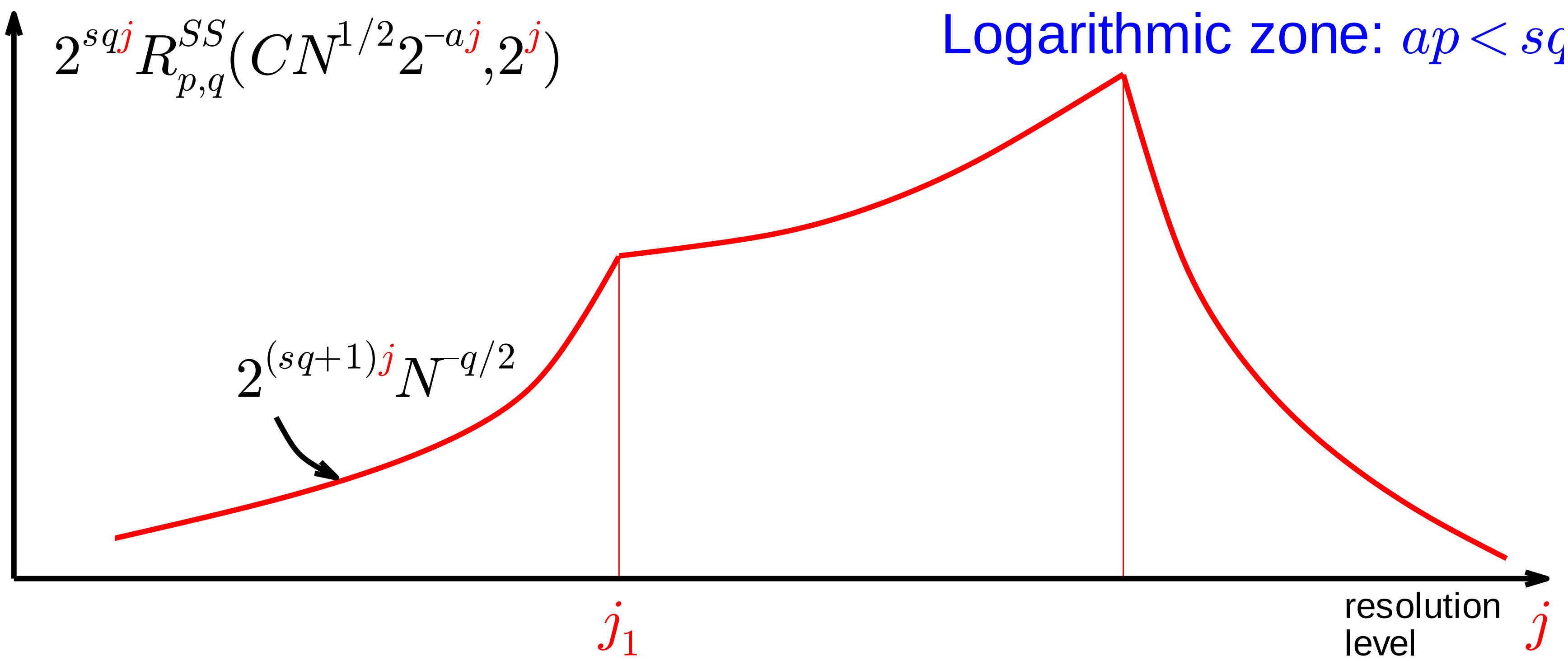
- Convert to noise level 1: $x_{jk} = N^{1/2} y_{jk}$
  has mean $\mu_{jk} = N^{1/2}\theta_{jk}$, variance 1.
- ℓ_p balls reduction:

$$\begin{aligned} \theta \in B^{\alpha}_{p,\infty}(C) \quad \Leftrightarrow \quad \|\theta_{\mathbf{j}}\|_p &\le C2^{-a\mathbf{j}} \quad \forall \mathbf{j} \\ \Leftrightarrow \quad 2^{-\mathbf{j}/p}\|\mu\|_p &\le \underbrace{CN^{1/2}2^{-(a+1/p)\mathbf{j}}}_{\eta_{\mathbf{j}}} \end{aligned}$$

- As $\mathbf{j} \nearrow$, $\quad \eta_{\mathbf{j}}$ crosses **dense, sparse, v. sparse** regions
- Singe sequence result (for **all regions**) $\Rightarrow$

$$\sup_{\|\theta_{\mathbf{j}}\| \le C2^{-a\mathbf{j}}} E\|\theta_{\mathbf{j}}{}^{EB} - \theta_{\mathbf{j}}\|_q^q \le c\big[R^{SS}_{p,q}(CN^{1/2}2^{-a\mathbf{j}}, 2^{\mathbf{j}}) + \cdots\big]$$

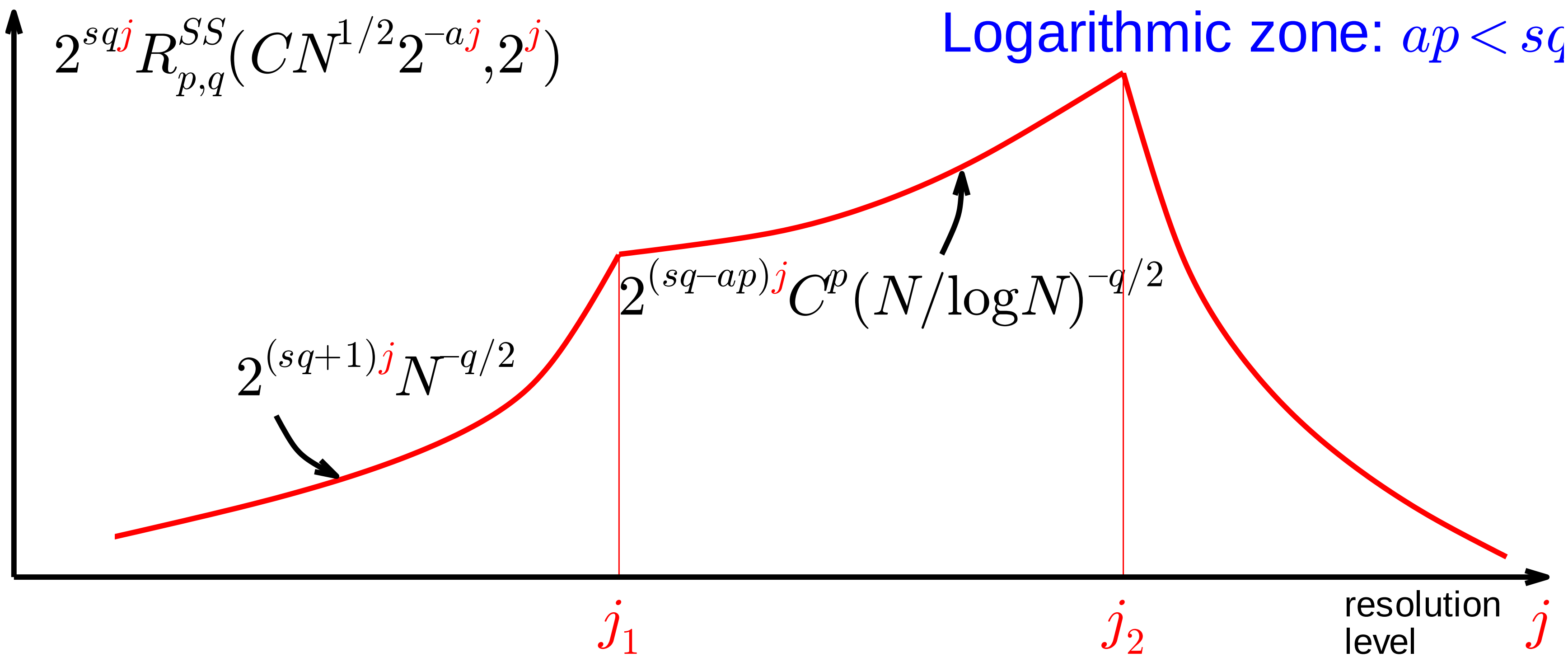
Single Sequence Phase Diagram transforms to:



dense

$$\eta_j = CN^{1/2}2^{-(a+1/p)j} > 1$$

Single Sequence Phase Diagram transforms to:



dense

$$\eta_j = CN^{1/2}2^{-(a+1/p)j} > 1$$

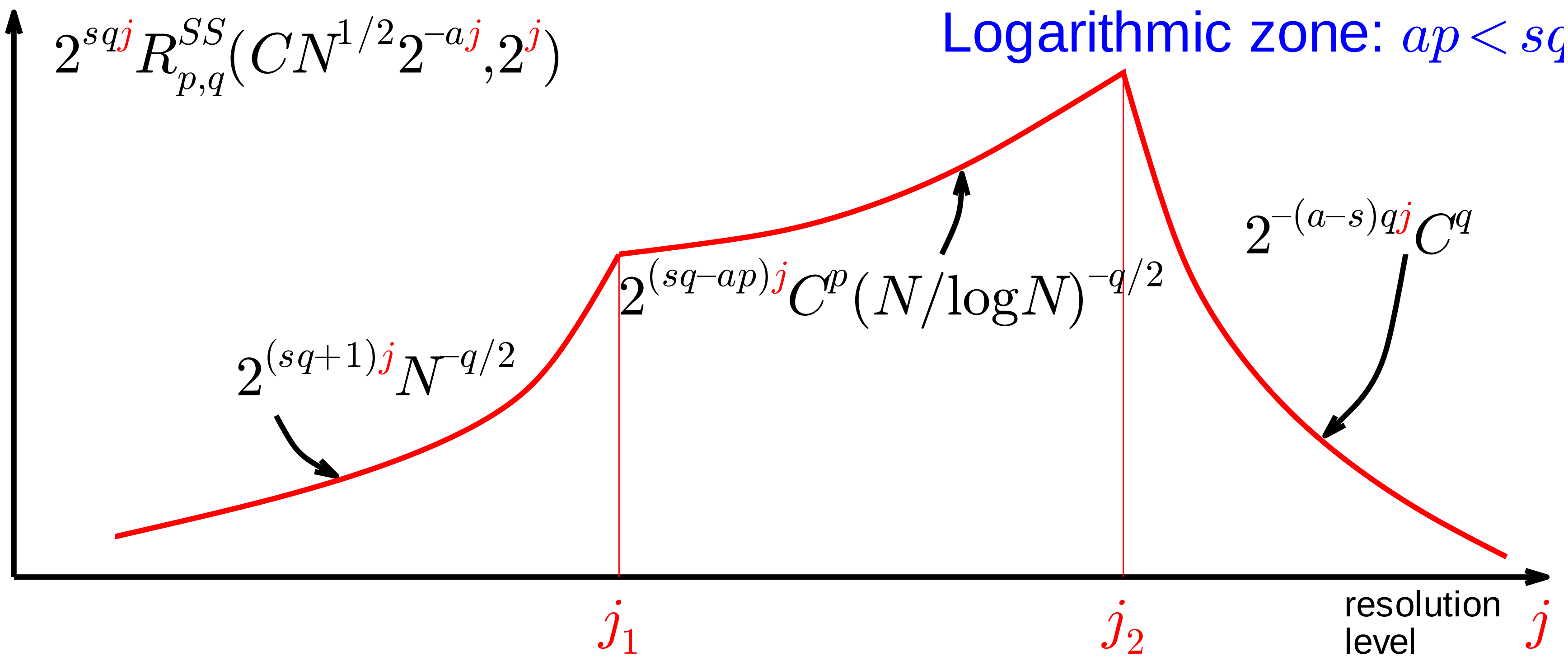
sparse

$$\eta_j < 1$$

and

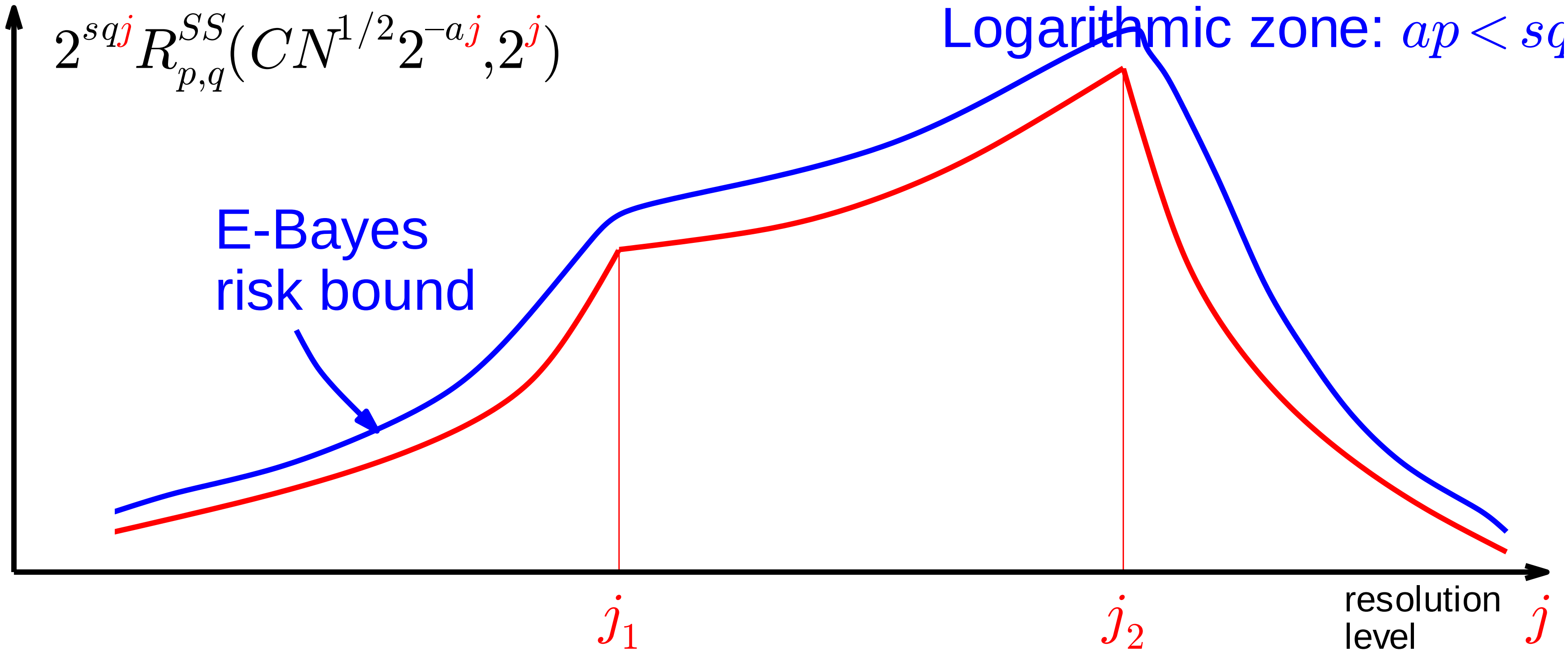
$$SN_j := CN^{1/2}\, 2^{-aj} > \sqrt{\log N}$$

Single Sequence Phase Diagram transforms to:



dense	sparse	v. sparse
$\eta_j = CN^{1/2} 2^{-(a+1/p)j} > 1$	$\eta_j < 1$	$\eta_j \ll 1$
	and	and
	$SN_j := CN^{1/2}\, 2^{-aj} > \sqrt{\log N}$	$SN_j < \sqrt{\log N}$

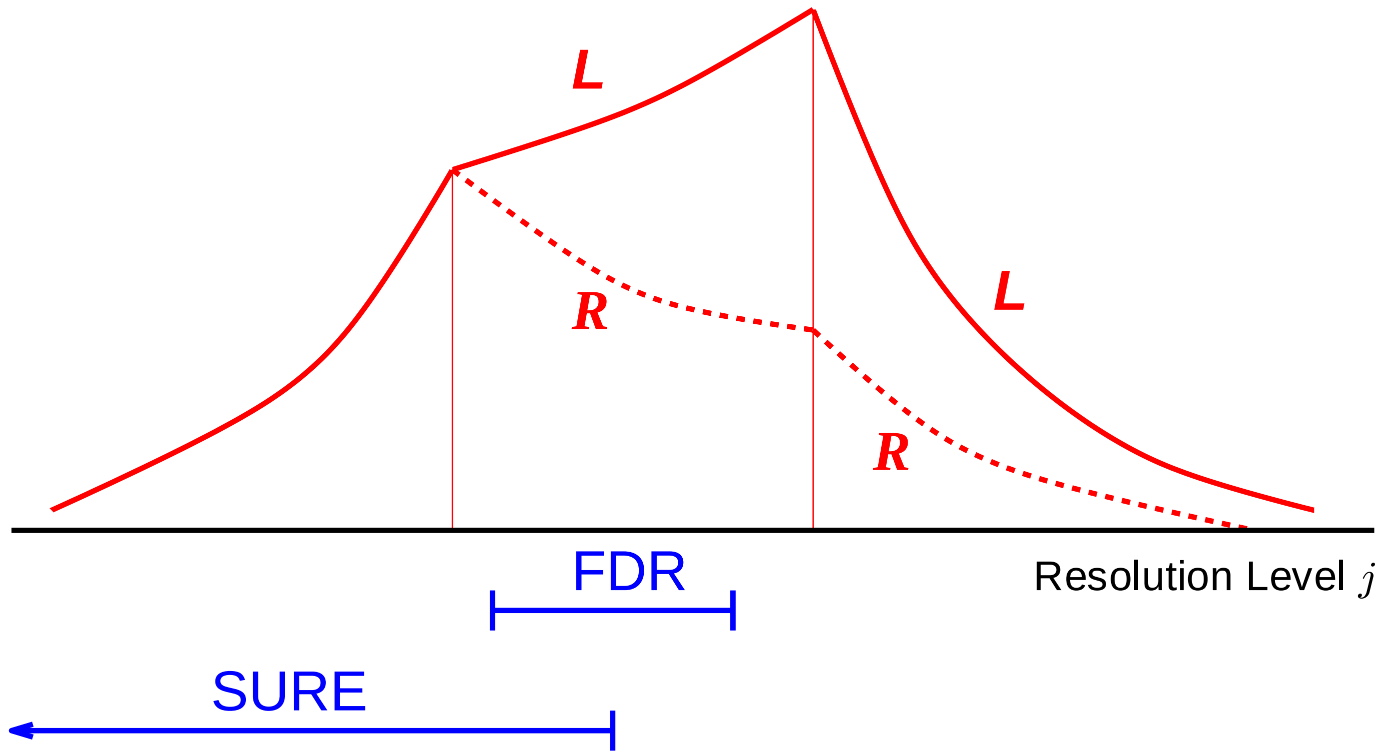
Single Sequence Phase Diagram transforms to:



dense | sparse | v. sparse

$\eta_j = CN^{1/2} 2^{-(a+1/p)j} > 1$ | $\eta_j < 1$ | $\eta_j \ll 1$

E-Bayes vs. FDR + SURE



- ℓ_p ball risk results for FDR, SURE each have limited range
- FDR (fine scales) combined with SURE (coarse scales) works for $\mathcal{R}$ (but not for $\mathcal{L}$?)

Where we're been

- Presence of **phase changes**

$$\textit{v.sparse} \quad \longrightarrow \quad \textit{sparse} \quad \longrightarrow \quad \textit{dense}$$

  challenges good threshold selection in sequence model.
- not all *a priori* reasonable methods are in fact satisfactory in adapting over the range of phases
- Some methods (**E-Bayes**, penalized least squares ...) can be shown to do so
- (adaptation to) phase changes in function estimation

$$\textit{Regular} \quad \longrightarrow \quad \textit{Critical} \quad \longrightarrow \quad \textit{Logarithmic}$$

  flows from phase changes in (growing) Gaussian sequence models

Take-away message

Minimax analysis, ℓ_p norms, function spaces, rates of convergence etc.,

– deepen understanding of a key methodological problem,

by yielding

- emergent phenomena *(phase changes)*
- a searching discipline of anaylsis *(adaptation)*
- aligned with practical considerations *(stability, reconstruction)*

and, not least, available software, at

- `www.stats.ox.ac.uk/~silverma/`

Three Talks

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

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 $n^{-1} X_n X_n^T$
   - PCA, CCA, MANOVA

FDR chooses large t in dense cases

- $k = 500$ of $n = 1000$ have $\mu_i = 3$, else 0 (10 reps).
- FDR at $q = .01, .05, .1$ chooses larger thresholds, with worse MSE, than E-Bayes at $a = .2, .5, 1$