Empirical Bayes: Methodologies and Asymptotic Theorems

1. Compound and empirical Bayes decision problems
2. Estimation of normal means
3. Nonparametric regression and the white noise with drift
4. Estimation of sums of random variables

Cun-Hui Zhang
Department of Statistics, Rutgers University

Thanks to: Zhe-Da, NSF, ...

Nonparametric Regression and the White Noise

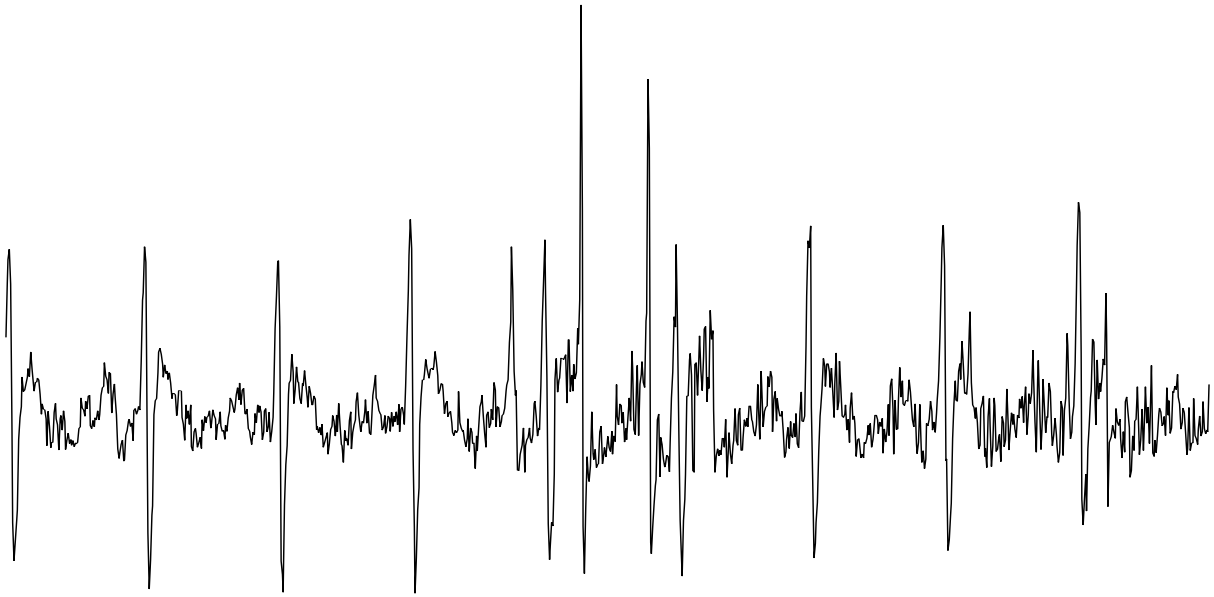
- Nonparametric regression and the white noise models
- Wavelets, smoothness, and sparsity
- Block empirical Bayes methods and wavelet denoising
- Theoretical properties of GEB methods
- Simulation with block wavelet methods

Acknowledgments: Donghui Zhang,
Ramprasath Lakshminarasimhan, Weihua Tang

Nonparametric regression and the white noise models

- Nonparametric Regression:

$$Y_i = f(t_i) + \varepsilon_i, \quad i \leq N.$$



- The white noise (Ibragimov-Khas'minskii, 81):

$$Y(t) = \int_0^t f(x)dx + \epsilon W(t), \quad \epsilon = \sigma/\sqrt{N}.$$

Problem: the estimation of f based on Y under the L_2 loss

$$R^{(\epsilon)}(\widehat{f}, f) \equiv E_f^{(\epsilon)} \int_0^1 \left\{\widehat{f}(t) - f(t)\right\}^2 dt.$$

- Orthonormal transformation (spline, wavelet, Fourier, etc.):

$$X_k \equiv \epsilon^{-1} \int \phi_k dY \sim N(\theta_k, 1)$$

for suitable basis functions ϕ_k, where

$$\theta_k = \epsilon^{-1}\beta_k, \quad \beta_k \equiv \int \phi_k f.$$

By Parseval, estimation of f is equivalent to that of $\beta \equiv \{\beta_k\}$:

$$R^{(\epsilon)}(\widehat{f}, f) = \sum_k E_\beta^{(\epsilon)} \left(\hat{\beta}_k - \beta_k\right)^2 = \epsilon^2 \sum_k E_\beta^{(\epsilon)} \left(\hat{\theta}_k - \theta_k\right)^2.$$

- Hopefully, the smoothness of f is reflected in the rate of $\beta_k \to 0$.
- Example: Sobolev balls for f with period 1:

$$S_\alpha(C) \equiv \left\{ f : \int_0^1 (f^{(\alpha)})^2 \le C^2 \right\}$$

with smoothness index α and radii C.

$\diamond$ For the Fourier basis functions $\phi_0(x) = 1$, $\phi_{2m-1}(x) = \cos(2\pi m x)$, and $\phi_{2m} = \sin(2\pi m x)$, $m = 1, 2, \ldots$, $f \in S_\alpha(C)$ iff

$$\sum_{m=1}^{\infty} m^{2\alpha}(\beta_{2m-1}^2 + \beta_{2m}^2) = \int_0^1 (f^{(\alpha)})^2 \le C^2.$$

Consequently, $\beta_k \sim k^{-2\alpha-1/2}$ in certain average sense.

- Risk for the NP regression:

$$R_N(\widehat{f}, f) \equiv N^{-1}\sum_{k=1}^{N} E_f\left\{\widehat{f}(t_i) - f(t_i)\right\}^2.$$

- Discrete orthonormal transformations:

$$X_k = N^{-1/2}\sum_{i=1}^{N}\phi_{ki}Y_i, \quad k \le N,$$

to map the NP regression data into a sequence to better capture the smoothness of f. For suitable $\phi_{ki} \approx \phi_k(t_i)$,

$$EX_k = \theta_k = N^{-1/2}\sum_i \phi_{ki} f(t_i) \approx \sqrt{N}\int \phi_k(t) f(t) \to 0.$$

◇ Equivalence:

$$R_N(\widehat{f}, f) = N^{-1}\sum_{k=1}^{N} E_f\left(\hat{\theta}_k - \theta_k\right)^2$$

◇ Normality: if $\varepsilon_i \sim N(0, \sigma^2)$, then $X_k \sim N(\theta_k, \sigma^2)$.

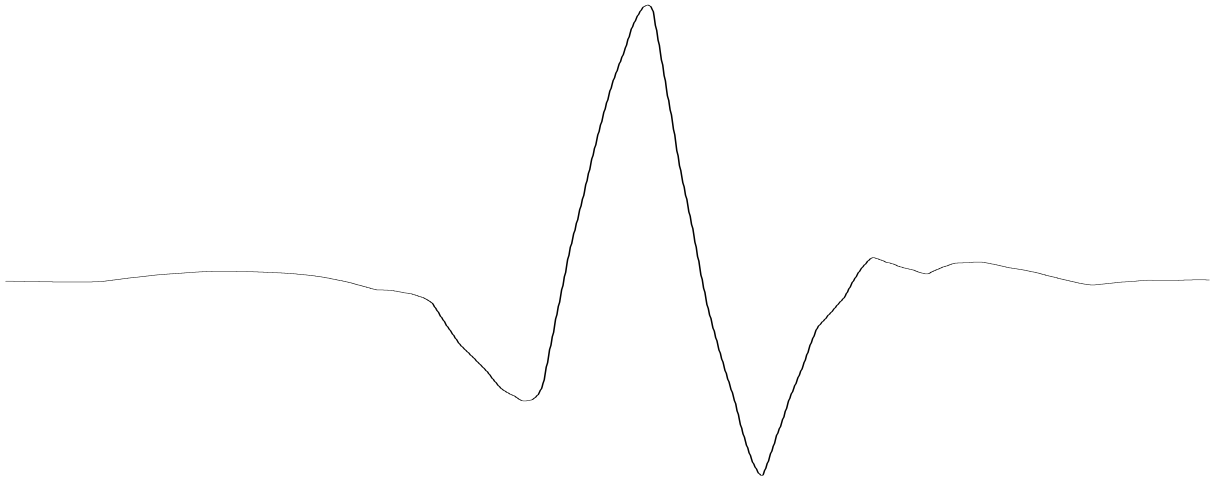
◇ FFT: $\phi_{2m-1,\ell} = \cos(2\pi m\ell/N)$, $\phi_{2m,\ell} = \sin(2\pi m\ell/N)$.

Wavelets, smoothness, and sparsity

- Wavelet basis:

$$\beta_{jk} = \int \phi_{jk} f, \quad k \leq 2^{j \vee 0}, j \geq -1,$$

where $\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k)$, $j \geq 0$, are "periodic" or "boundary adjusted" dilation and translation of a mother wavelet, and $\phi_{-1,0}$ is the father wavelet.



• Smoothness of f is reflected in the rate at which $\beta_{[j]} \equiv \{\beta_{jk}, k \leq 2^j\}$ converges to 0 under certain norm, as in the case of Fourier basis for functions in Sobolev balls.

• Sparsity of wavelet coefficients: the spatial inhomogeneity of f is often reflected in the sparsity of its wavelet coefficients $\beta_{[j]}$ in individual resolution levels, not necessarily in the overall smoothness.

Thus, in wavelet denoising, adaptation to the spatial inhomogeneity of f is achieved via (minimax) adaptive estimation for sparse wavelet coefficients.

◇ Chui (92), Daubechies (92), Donoho-Johnstone (94), Donoho *et al* (95), Härdle *et al* (98), ...

• Besov balls: $B^{\alpha}_{p,q}(C) \equiv \{\beta : \|\beta\|^{\alpha}_{p,q} \leq C\}$, i.e.

$$B^{\alpha}_{p,q}(C) \equiv \left\{\beta : \sum_j \left(2^{j(\alpha+1/2-1/p)} \|\beta_{[j]}\|_{p,2^j}\right)^q \leq C^q\right\},$$

where $\beta \equiv \{\beta_{jk}\}$. Here, α is the degree of smoothness, and (p, q) are shape parameters. A sequence $\beta = \{\beta_{jk}\}$ is in $B^{\alpha}_{p,q}(C)$ iff

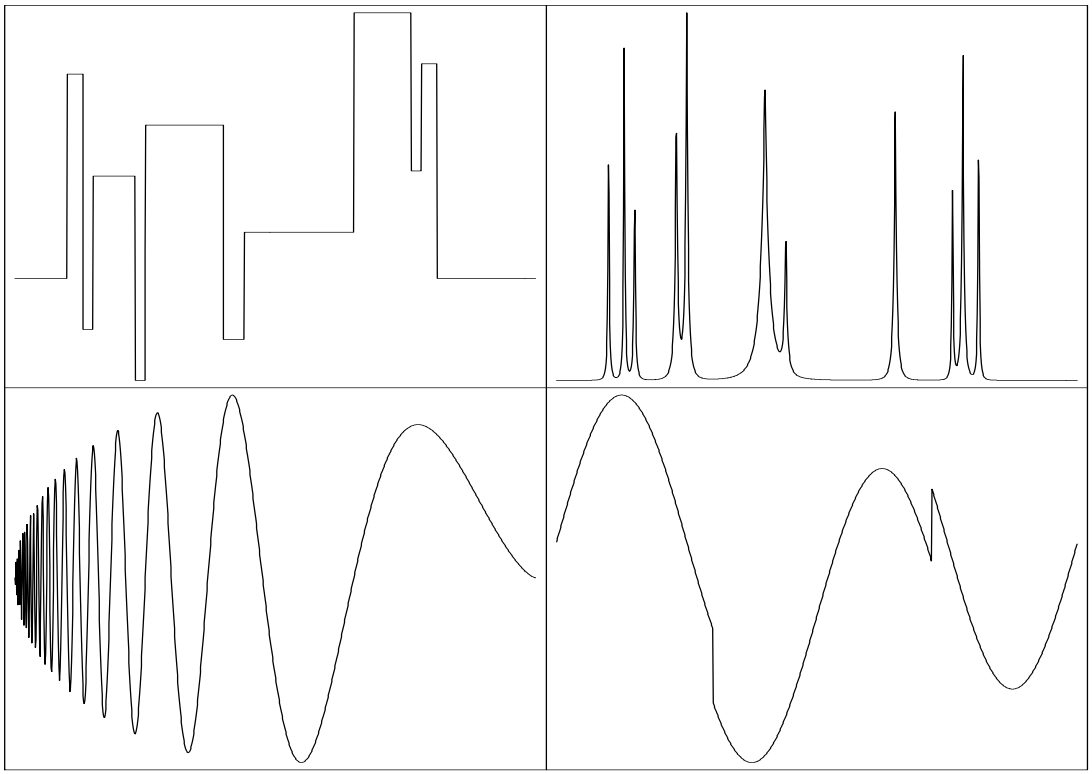
$$\beta_{[j]} \in \Theta_{p,2^j}(2^{-j(\alpha+1/2)} C_j), \quad \sum_j C_j^q \leq C^q.$$

• Example: Let $\mathcal{F}_{d,m}(C)$ be the collection of all piecewise polynomials f of degree d in $[0, 1]$ with at most m pieces and $\|f\|_\infty \leq C$. Let $\int t^j \phi(t) = 0$ for $j \leq d$. Then,

$$\|\beta_{[j]}\|_{p,2^j} \leq 2^{-j/2} m^{1/p} C M_0, \quad \|\beta\|^{\alpha}_{p,q} < \infty \Leftrightarrow \alpha < 1/p,$$

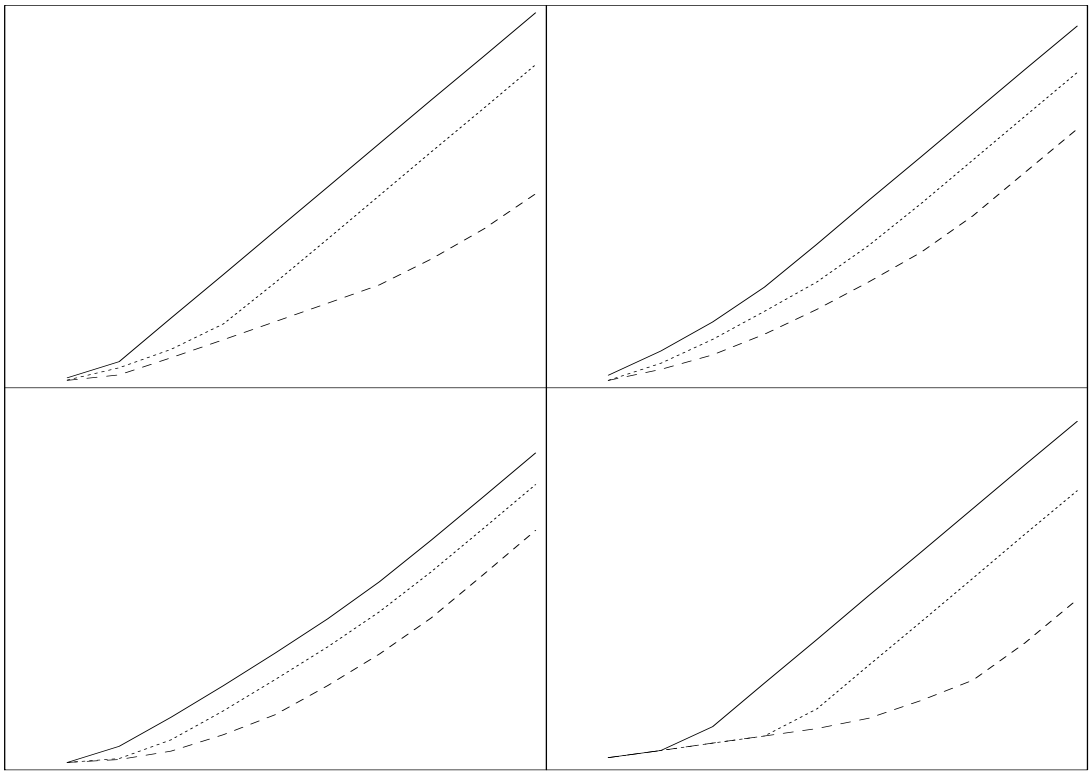
($\alpha = 1/p$ for $q = \infty$).

Examples of (artificial, sparse) signals:



Clockwise from top left: blocks, bumps, heavisine, doppler

The logarithm of Besov norms vs $\alpha = (1:10)/2$, $p \in \{1, 1/2, 1/5\}$, $q = \infty$, max ≈ 35



Clockwise from top left: blocks, bumps, heavisine, doppler

Block empirical Bayes methods and wavelet denoising

- Block empirical Bayes estimators for the white noise

◇ Data: $X_k \sim N(\theta_k, 1)$, $\theta_k = \beta_k/\epsilon \to 0$

◇ Problem: estimate $\{\beta_k\}$ under ℓ_2 loss.

◇ Block estimators: estimate $\theta_{[j]}$ based on $X_{[j]}$, where $[j] = (k_{j-1}, k_j]$ for suitable $k_j \uparrow$.

- Implementation in the NP regression model

◇ Estimate the noise level σ using the last block, $X_{[m]}$ with $k_m = N$, and treat $\tilde{X}_k \equiv X_k/\hat{\sigma}$ as data in the white noise model. For example,

$$\hat{\sigma} = \mathrm{MAD}(X_{[m]}).$$

• Wavelet denoising: natural blocks, $\beta_{[j]} = \{\beta_{jk}, k \le 2^j\}$.

◇ Goal: finding exactly adaptive (simultaneously asymptotically) minimax estimators $\hat{\beta}^{(\epsilon)}$,

$$\sup_{\beta \in B} R^{(\epsilon)}(\hat{\beta}^{(\epsilon)}, \beta) \le (1 + o(1))\mathcal{R}^{(\epsilon)}(B), \quad \forall\, B \in \mathcal{B}$$

for a large collection $\mathcal{B}$ of sets B of β, where $\mathcal{R}^{(\epsilon)}(B)$ is the minimax risk

$$\mathcal{R}^{(\epsilon)}(B) \equiv \inf_{\hat{\beta}} \sup_{\beta \in B} R^{(\epsilon)}(\hat{\beta}, \beta).$$

◇ The case of $\mathcal{B} = \{$ all Besov balls $B^{\alpha}_{p,q}(C)\}$ is of particular interest, since adaptive minimaxity for $p < 2$ implies spatial adaptation for the estimation of f.

• Efromovich-Pinsker (84, 86, modified block JS): *exactly adaptive minimax* in Sobolev balls (all α, $p = 2$).

• Johnstone-Silverman (04, EB posterior median): *rate* adaptive minimax threshold methods in Besov balls (all α and p).

• Donoho-Johnstone (95, SureShrink): exactly adaptive to the ideal threshold risk for certain (α, p) with $p < 2$.

• Abramovich *et al* (04, FDR): exactly adaptive for sparse signals.

• Additional references (adaptive NP methods, minimaxity, ...): Breiman *et al* (83), Stone (84), Efromovich (85), Friedman (91), Golubev (92), Johnstone *et al* (92), Foster-George (94), Hall-Patil (95,96), Donoho *et al* (96), Brown *et al* (97), Juditsky (97), Lepski *et al* (97), Härdle *et al* (98), Hall *et al* (98,99), Barron *et al* (99), Cai (99), Cavalier-Tsybakov (01), ...

• Minimax convergence rates in Besov balls (DJ, 98):

$$\mathcal{R}^{(\epsilon)}\big(B^{\alpha}_{p,q}(C)\big) \asymp \epsilon^{2\alpha/(\alpha+1/2)} C^{1/(\alpha+1/2)} = \epsilon^2 (C/\epsilon)^{1/(\alpha+1/2)}$$

as $C/\epsilon \to \infty$, where $\mathcal{R}^{(\epsilon)}(B)$ is the minimax risk.

◇ For certain C_j with $\sum_j C_j^q \le C^q$, necessarily $C_j \le C$,

$$\beta \in B^{\alpha}_{p,q}(C) \Rightarrow \theta_{[j]} \in \Theta_{p,2^j}(2^{-j(\alpha+1/2)} C_j/\epsilon).$$

◇ Recall that $R^{(\epsilon)}(\hat\beta, \beta) = \epsilon^2 \sum_{j,k} E^{(\epsilon)}_{\beta}(\hat\theta_{jk} - \theta_{jk})^2$.

◇ Dense or large signals: $2^{-j(\alpha+1/2)} C/\epsilon \asymp 1$ or $2^j \asymp (C/\epsilon)^{1/(\alpha+1/2)}$ or smaller j, totally $O(1)(C/\epsilon)^{1/(\alpha+1/2)}$ parameters θ_{jk} to estimate.

◇ Small signals: $(C/\epsilon)^{1/(\alpha+1/2)} = o(2^j)$ and $p \ge 2$

◇ Sparse signals: $(C/\epsilon)^{1/(\alpha+1/2)} = o(2^j)$ and $p < 2$

- Adaptive minimax estimation in Besov balls:

$$\mathcal{R}^{(\epsilon)}\left(B^{\alpha}_{p,q}(C)\right) \leq \sup_{\{C_j\}} \epsilon^2 \sum_j \mathcal{R}_{2^j}\left(\Theta_{p,2^j}\left(2^{-j(\alpha+1/2)}C_j/\epsilon\right)\right)$$

can be achieved if an estimator is

$\diamond$ boundedly minimax for dense or large signals, totally $o(1)(C/\epsilon)^{1/(\alpha+1/2)}$ parameters

$\diamond$ exactly adaptive minimax for dense signals, totally $O(1)(C/\epsilon)^{1/(\alpha+1/2)}$ parameters

$\diamond$ rate adaptive for sparse signals, for potentially infinitely many parameters.

- Cut-off at $2^j \asymp (\epsilon/C)^{1/(\alpha+1/2)}$ yields optimal minimax rates for $p \geq 2$ and known α.

- The (Block) James-Stein estimator are exactly (rate) adaptive minimax in Besov balls for $p = 2$ ($p > 2$).

- The (Block) SureShrink is rate adaptive minimax in Besov balls for certain α and $p \leq 2$.

- The (Block) EB posterior median is rate adaptive minimax in all Besov balls.

- It is not clear if the (block) FDR threshold method are rate adaptive minimax, since its properties for dense and large signals are unclear.

- We discuss properties of (block) GEB estimators in detail.

Theoretical properties of GEB methods

- Ideal adaptation
- Adaptive minimaxity
- Spacial adaptation
- Dominance over certain other methods
- Universal super-efficiency

- In what follows, let $\hat{\beta}^{(\epsilon)}$ be the (block) HGEB estimator.
- Ideal adaptation: adaptation to ideal/Bayes risks
  - $\diamond$ Ideal risk: with $\mathcal{D}^s$ being the set of all separable $\hat{\beta}_{jk} = t_j(X_{jk})$,

$$R^{(\epsilon,*)}(\beta) \equiv \min_{\beta \in \mathcal{D}^s} R^{(\epsilon)}\big(\hat{\beta}, \beta\big).$$

  - $\diamond$ $R^{(\epsilon,*)}(\beta)$ is the risk of an ideal Bayes estimator.
  - $\diamond$ Oracle inequalities in individual blocks yield

$$\sup_{\beta \in B^{\alpha}_{p,q}(C)} \Big\{ R^{(\epsilon)}\big(\hat{\beta}^{(\epsilon)}, \beta\big) - R^{(\epsilon,*)}(\beta) \Big\} \leq o(\epsilon^{\gamma}) \mathcal{R}^{(\epsilon)}\big(B^{\alpha}_{p,q}(C)\big)$$

for certain $\gamma > 0$. In this sense, the HGEB achieves ideal adaptation.

• Adaptive minimaxity: the HGEB is exactly adaptive minimax in all Besov balls

$$\sup_{\beta \in B^{\alpha}_{p,q}(C)} R^{(\epsilon)}\big(\hat{\beta}^{(\epsilon)}, \beta\big) \leq (1 + o(\epsilon^{\gamma}))\mathcal{R}^{(\epsilon)}\big(B^{\alpha}_{p,q}(C)\big).$$

◇ The minimax equivalence between the compound and Bayes estimation problems yields the minimax theorem:

$$\sup_{\beta \in B^{\alpha}_{p,q}(C)} R^{(\epsilon,*)}(\beta) \leq (1 + o(\epsilon^{\gamma}))\mathcal{R}^{(\epsilon)}\big(B^{\alpha}_{p,q}(C)\big).$$

◇ Ideal adaptation + minimax th. $\Rightarrow$ exactly adapt. minimaxity.

◇ Extension of Efromovich-Pinsker (84, 86).

• Spacial adaptation: exactly adaptive minimaxity in all Besov balls (i.e. for small p) $\Rightarrow$ adaptation to spatial inhomogeneity of f.

• Dominance over certain other methods: if

$$R^{(\epsilon)}\big(\tilde{\beta}^{(\epsilon)},\beta\big) \geq (1+o(1)) \inf_{\hat{\beta}\in\mathcal{D}_0} R^{(\epsilon)}\big(\hat{\beta},\beta\big)$$

for a smaller class (e.g. separable threshold estimators) $\mathcal{D}_0 \subseteq \mathcal{D}^s$, then

$$\sup_{\beta\in B} R^{(\epsilon)}\big(\tilde{\beta}^{(\epsilon)},\beta\big) \geq (1+o(1)) \sup_{\beta\in B} R^{(\epsilon)}\big(\hat{\beta}^{(\epsilon)},\beta\big)$$

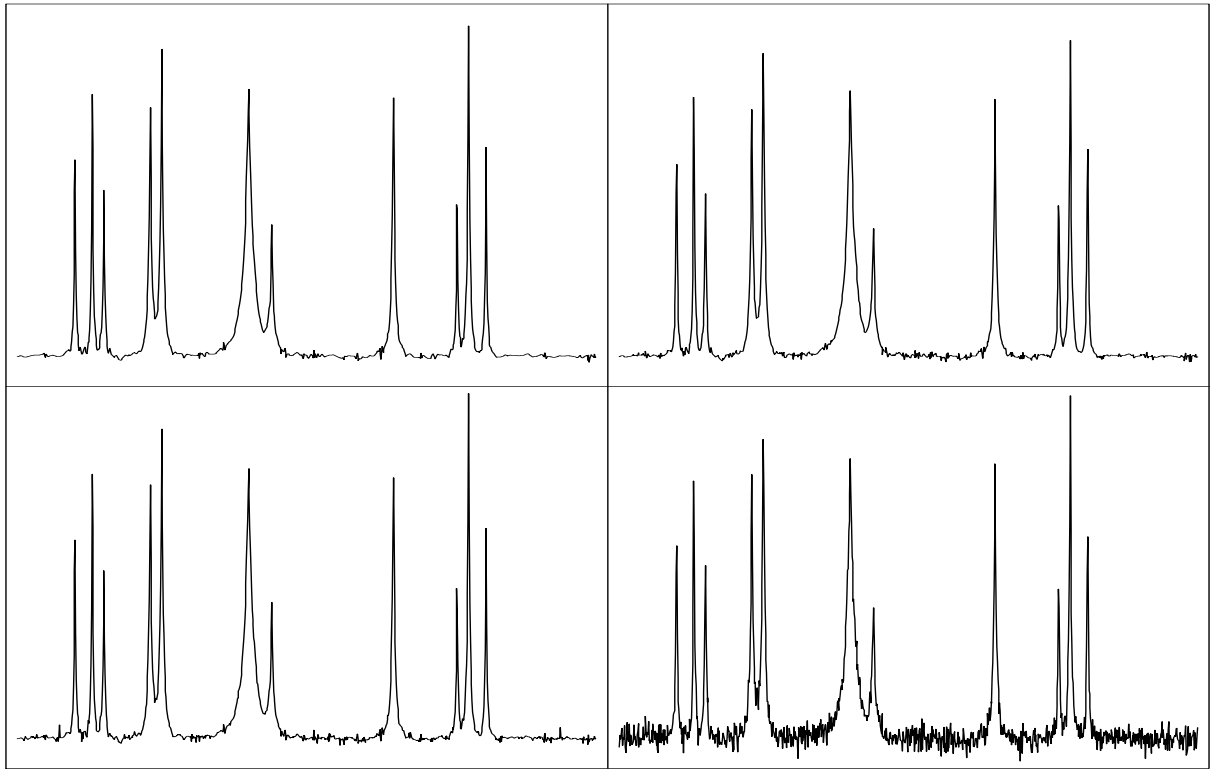
for all B, such that $B \subseteq B^{\alpha}_{p,q}(C)$ and $\mathcal{R}^{(\epsilon)}(B) \geq c_0\epsilon^{2\alpha/(\alpha+1/2)}$.

• Universal super-efficiency: for $q<\infty$ and all compact sets B under $\|\cdot\|^{\alpha}_{p,q}$

$$\sup_{\beta\in B} R^{(\epsilon)}\big(\hat{\beta}^{(\epsilon)},\beta\big) = o(1)\mathcal{R}^{(\epsilon)}(B^{\alpha}_{p,q}(C)).$$

Simulation with block wavelet methods

signal = "bumps", $N = 4096$, $SNR = 7$



Clockwise from top left: ebayesthresh, sureshrink, fdr, rgeb

Table 2.1. Average of 100 simulated SSE
"heavisine", $N = 1024$, $SNR = 7$

j		6	7	8	9	10	11	total
n_j	16	16	32	64	128	256	512	1024
SS sig	49	0	0	0	0	0	0	49
JS	16	14	24	25	11	9	10	109
Sure	16	12	12	12	10	11	12	86
EBPM	16	11	9	8	8	9	10	71
FDR	16	12	10	10	8	9	10	77
RGEB	16	15	16	19	23	31	39	160
HGEB	16	15	15	12	8	9	10	84
PGEB	16	15	14	14	13	17	21	111

Table 2.2. "bumps", $N = 1024$, $SNR = 7$

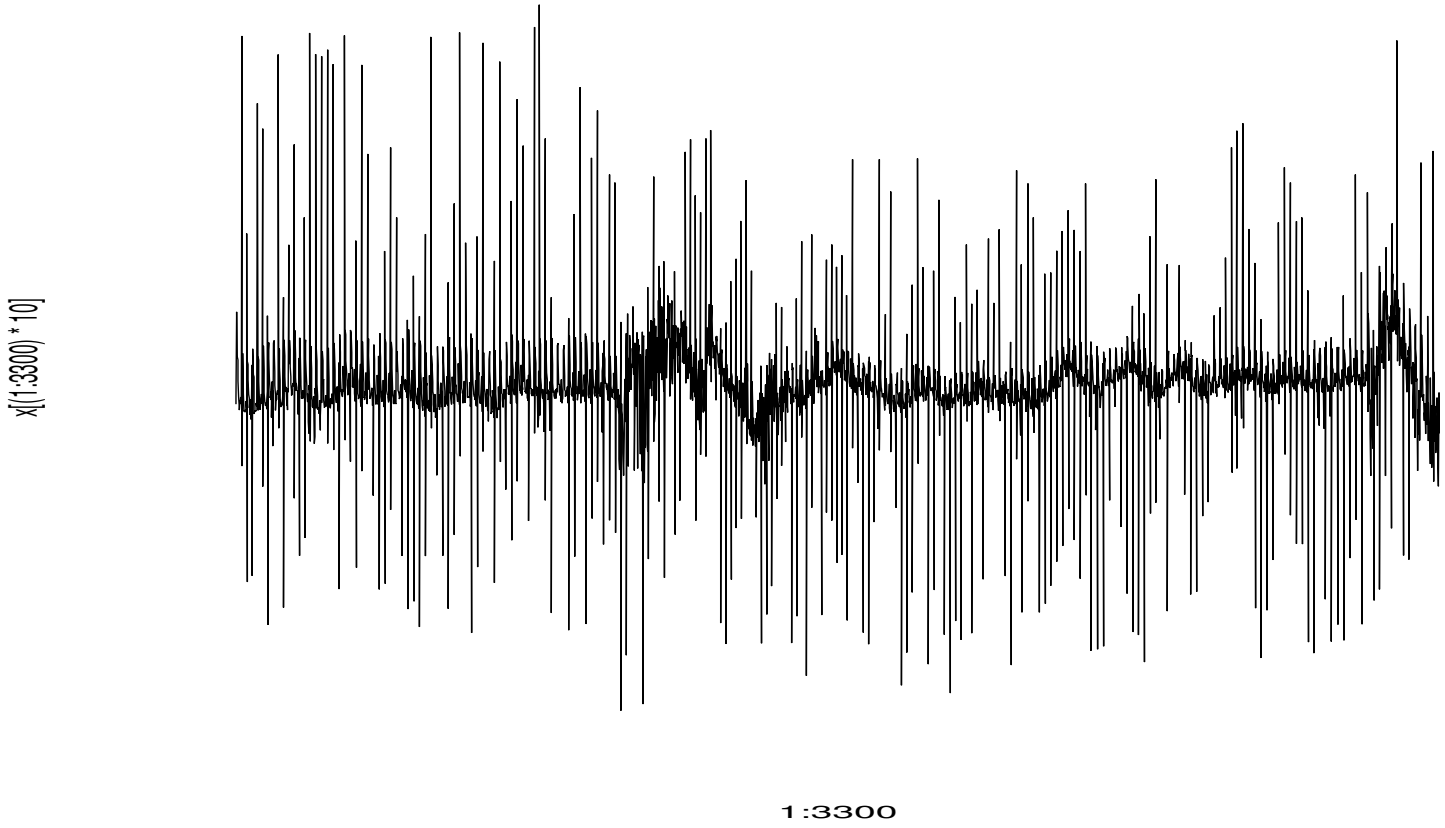
j		6	7	8	9	10	11	total
n_j	16	16	32	64	128	256	512	1024
SS sig	20	9	4	9	4	2	1	49
JS	16	15	33	63	125	233	264	750
Sure	16	24	30	49	78	101	92	391
EBPM	16	17	30	47	65	77	79	331
FDR	16	18	36	45	70	88	89	362
RGEB	16	17	35	46	68	88	94	364
HGEB	16	17	35	46	68	87	90	360
PGEB	16	17	35	46	68	87	90	360

Table 2.3. Four "bumps", $N = 4096$, $SNR = 7$

j		6	7	8	9	10	11	total
n_j	64	64	128	256	512	1024	2048	4096
SS sig	20	9	4	9	4	2	1	49
JS	64	64	128	251	497	909	1040	2954
Sure	64	68	116	189	305	385	343	1470
EBPM	64	75	116	180	258	301	299	1294
FDR	64	71	132	181	290	354	341	1433
RGEB	64	62	118	158	242	305	313	1262
HGEB	64	62	118	158	246	311	325	1284
PGEB	64	62	118	158	248	321	338	1310

Further research

An example of real data:



Conclusions

- Adaptation to spatial inhomogeneity of signals can be achieved through adaptive minimax estimation for sparse signals
- Hybrid GEB estimators possess a number of optimality properties, including exactly adaptive minimaxity in all Besov balls
- Our simulation study of the GEB and threshold estimators provides further evidence for the validity of the asymptotic theory