Bayesian wavelet estimators in nonparametric regression

Lecture 4

Natalia Bochkina, University of Edinburgh

Lecture 4. Wavelet estimators: simultaneous local and global optimality

1. Separable and non-separable function estimators

2. When simultaneous local and global optimality is possible

3. Bayesian wavelet estimator that is locally and globally optimal

4. Conclusions and open questions

Quick introduction

Wavelet nonparametric regression:

$$y_i = f(i/n) + \epsilon_i, \quad i = 1, \ldots, n,$$

assuming $\mathbb{E}\epsilon_i = 0$, $\mathbb{E}\epsilon_i^2 < \infty$ and ϵ_i are independent.

Aim: estimate f.

Model function via its orthonormal wavelet decomposition

$$f(x) = \sum_{k=0}^{2^L-1} \theta_k \phi_{Lk}(x) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk} \psi_{jk}(x)$$

Apply orthonormal discrete wavelet transform:

$$d_{jk} = w_{jk} + \varepsilon_{jk}, \quad c_{Lk} = u_{Lk} + \varepsilon_{L-1,k}.$$

$j = L, L+1, \ldots, J-1$, $k = 0, 1, \ldots, 2^j - 1$, $n = 2^J$.

If $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ then $\varepsilon_{jk} \sim \mathcal{N}(0, \sigma^2)$ iid.

Rescaling discrete wavelet transform

Discrete wavelet transform:

$$d_{jk} = w_{jk} + \varepsilon_{jk}, \quad c_{Lk} = u_{Lk} + \varepsilon_{L-1,k}.$$

Relationship to "continuous" wavelet coefficients θ: $\tilde{\theta}_{jk} \stackrel{def}{=} \frac{w_{jk}}{\sqrt{n}} \approx \theta_{jk}$.

To esti mate f, need to estimate w_{jk} and u_{Lk}.

Rescale:

$$y_{jk} \stackrel{def}{=} d_{jk}/\sqrt{n}, \quad y_k \stackrel{def}{=} c_{L,k}/\sqrt{n},$$

and for independent Gaussian noise,

$$y_{jk} \mid \tilde{\theta}_{jk} \sim \mathcal{N}(\tilde{\theta}_{jk}, \sigma^2/n), \quad y_k \mid \tilde{\theta}_k \sim \mathcal{N}(\tilde{\theta}_k, \sigma^2/n),$$

independently.

Aims

1. Ideal aim: given a loss function, characterise the set of Bayesian models that result in adaptive wavelet estimators optimal with respect to this loss.

Considered loss functions:

- local error $d(f, \hat{f}_n) = |f(x_0) - \hat{f}_n(x_0)|^u$, $u \in [1, \infty)$
- global error $d(f, \hat{f}_n) = ||f - \hat{f}_n||_u$, $u \in [1, \infty]$.

For wavelet estimators: minimax optimality over Besov spaces $B^r_{p,q}$.

2. Next step is to look at the intersection of these models, i.e. are there Bayesian models that are simultaneously locally and globally optimal?

Besov space $B^r_{p,q}$

Besov sequence norm of wavelet coefficients θ with $r > 0$, $1 \leq p, q \leq \infty$ is defined by

$$
\begin{aligned}
||\theta||_{b^r_{p,q}} &= ||\theta_k||_p + \left[\sum_{j=L}^{\infty} 2^{qj(r+1/2-1/p)} ||\theta_j||_p^q\right]^{1/q}, \quad \text{if} \quad q < \infty, \\
||\theta||_{b^r_{p,q}} &= ||\theta_k||_p + \sup_{L \leq j < \infty} [2^{j(r+1/2-1/p)} ||\theta_j||_p], \quad \text{if} \quad q = \infty,
\end{aligned}
$$

and the Besov sequence space $b^r_{p,q}(A) = \{\theta : \quad ||\theta||_{b^r_{p,q}} \leq A\}$.

If regularity of wavelets s is such that $s > r > 0$, then Besov sequence norm is equivalent to Besov space norm (Donoho and Johnstone, 1998).

In particular, if $\theta \in b^r_{p,q}(A)$, then $||\theta_j||_p \leq A 2^{-j(r+1/2-1/p)}$.

Minimax optimal rates for Besov spaces

Local minimax rate and adaptivity

For Besov spaces $B^r_{p,q}$ with $r > 1/p$ and $u \in [1, \infty)$, the local minimax rate is given by

$$\inf_{\tilde{f}_n} \sup_{f \in B^r_{p,q}} \mathbb{E}|\tilde{f}_n(t_0) - f(t_0)|^u \asymp n^{-\frac{u(r-1/p)}{2(r-1/p)+1}},$$

however, this rate cannot be achieved by an adaptive estimator (e.g. Lepski(1990) for Hölder spaces and Brown and Low (1996)), and the best possible local rate an adaptive estimator can achieve is given by

$$\inf_{\tilde{f}_n} \sup_{f \in B^r_{p,q}} \mathbb{E}|\tilde{f}_n(t_0) - f(t_0)|^u \asymp \left(\frac{n}{\log n}\right)^{-\frac{u(r-1/p)}{2(r-1/p)+1}}.$$

Price for adaptivity: $[\log n]^{\frac{u(r-1/p)}{2(r-1/p)+1}}$.

Global minimax rate for Besov spaces

For Besov spaces $B^r_{p,q}$ and $u \in [1, \infty)$, global minimax rate is given by

$$\inf_{\tilde{f}_n} \sup_{f \in B^r_{p,q}} \mathbb{E}||\tilde{f}_n - f||_u^u \asymp \Lambda_{r,p}(n)$$

where

$$\Lambda_{r,p}(n) = \begin{cases} n^{-\frac{ur}{2r+1}}, & \text{if} \quad u < p(2r+1), \\ \left(\frac{n}{\log n}\right)^{-\frac{u(r-1/p)+1}{2(r-1/p)+1}} \log n, & \text{if} \quad u = p(2r+1), \\ \left(\frac{n}{\log n}\right)^{-\frac{u(r-1/p)+1}{2(r-1/p)+1}}, & \text{if} \quad u > p(2r+1) \end{cases}$$

(Donoho et al. (1995, 1996) and Delyon and Juditsky (1996)).

Can be achieved for adaptive estimators.

What kind of adaptive estimators achieve this rate?

Separability

Sequence model:

$$y_{jk} = \theta_{jk} + \varepsilon_{jk}, \quad j = L, \ldots, \infty, \quad k = 0, \ldots, 2^j - 1,$$

where $\varepsilon_{jk} \sim \mathcal{N}(0, \sigma^2)$ independent.

Definition 1. *Estimator $\delta = (\delta_{jk})$ is separable if for all (j, k) δ_{jk} depends only on y_{jk} and not any other ys.*

Non-adaptive estimation

Donoho and Johnstone (1998) showed that

- Bayes minimax estimators for a Besov body $b^r_{pq}(A)$ and L^2 loss are separable
- The optimal separable estimators are asymptotically minimax when $p \leq q$ and are within a constant factor of minimax when $p > q$

Depend on parameters of $b^r_{pq}(A)$.

Adaptation for separable estimators (Cai, 2008)

Denote a set of separable estimators by $\mathcal{E}_n$.

Theorem 1. *(Theorem 1 of Cai, 2008). If $\delta_n \in \mathcal{E}_n$ attains the optimal rate of convergence over a Besov body $b^r_{pq}(A)$, then it must attain the exact same rate at every point:*

$$0 < \liminf_{n\to\infty} n^{2r/(2r+1)}\mathbb{E}||\delta_n - \theta||_2^2 \leq \limsup_{n\to\infty} n^{2r/(2r+1)}\mathbb{E}||\delta_n - \theta||_2^2 < \infty$$

for every $\theta \in b^r_{pq}(A)$.

Hence, **separable estimators are not rate adaptive**.

For an estimator to be rate adaptive, it has to pool information across observations.

Definition 2. *Estimator δ is called superefficient at a fixed point $\theta \in b^r_{pq}(A)$ if $n^{2r/(2r+1)}\mathbb{E}||\delta - \theta||_2^2 \to 0$ as $n \to \infty$.*

For an estimator δ_n to be adaptive, i.e. to change the convergence rate on a "smoother" subset of b^r_{pq}, it has to be superefficient on this set.

Minimum cost for adaptation for separable estimators

Theorem 2. *(Theorem 2 of Cai, 2008). Suppose $r_1 > r_2$. If a separable rule δ_n attains a rate of $n^{-\rho}$ over $b^{r_1}_{p_1 q_1}(A_1)$ with $\rho > 2r_2/(1+2r_2)$, in particular, if δ_n is rate-optimal over $b^{r_1}_{p_1 q_1}(A_1)$, then*

$$\liminf_{n\to\infty} \left(\frac{n}{\log n}\right)^{2r_2/(1+2r_2)} \sup_{\theta \in b^{r2}_{p_2 q_2}(A_2)} \mathbb{E}||\delta_n \theta||_2^2 > 0.$$

That is, the rate of convergence over $b^{r_2}_{p_2 q_2}(A_2)$ cannot be faster than $\left(\frac{n}{\log n}\right)^{-2r_2/(1+2r_2)}$.

Hence the minimum cost for adaptation for the separable rules over $b^r_{pq}(A)$ is at least a logarithmic factor $\log n^{2r/(1+2r)}$.

The thresholding estimator with universal threshold achieves this rate adaptively across a wide range of b^r_{pq} (Donoho & Johnstone, 1994), hence this bound is sharp.

Information pooling and adaptability

Theorem 3. *(Theorem 3 of Cai, 2008). Suppose* $r > 0$ *and let* $\delta = (\delta_{jk})$ *be an estimator that depends on at most* $h_n = o((\log n)^{2r/(1+2r)})$ *observations. Let* $\theta^\star \in b^r_{pq}(A)$*. If*

$$\lim_{n\to\infty}\inf n^\rho \mathbb{E}||\delta - \theta^\star||_2^2 < \infty$$

for some $\rho > 2r/(1+2r)$*, then*

$$\lim_{n\to\infty}\inf n^{2r/(1+2r)} \frac{h_n}{(\log n)^{2r/(1+2r)}} \sup_{\theta \in b^r_{pq}(A)} \mathbb{E}||\delta - \theta||_2^2 > 0.$$

In particular,

$$\lim_{n\to\infty}\inf n^{2r/(1+2r)} \sup_{\theta \in b^r_{pq}(A)} \mathbb{E}||\delta - \theta||_2^2 = \infty.$$

Therefore, in order to achieve adaptability, the information pooling index h_n should be at least of order $(\log n)^{2r/(1+2r)}$.

Block James-Stein estimator, which is adaptive and attains the optimal global rate for a wide range of b^r_{pq}, has $h_n = \log n$.

Block James-Stein estimator

Divide each resolution level $j < J$ into nonoverlapping blocks of approximate length $\ell = \log n$.

Denote (jb) the b-th block at level j and $S^2_{(jb)} = \sum_{k \in (jb)} y^2_{jk}$.

Let $\lambda^* = 4.50524$ be the root of the equation $\lambda - \log \lambda - 3 = 0$. The BlockJS estimator $\hat{f}_{BJS}$ has the following non-zero wavelet coefficients $\hat{\theta}_{jk}$:

$$\hat{\theta}_{jk} = \left(1 - \frac{\lambda^* \ell \sigma^2 n^{-1}}{S^2_{(jb)}}\right)_+ y_{jk}, \quad k \in (jb), \quad j < J.$$

Then, Cai(1999) showed that $\hat{f}_{BJS}$ achieves minimax global rate for $u = 2$ over $B^r_{pq}(A)$ for all $r \in (0, s)$, $p \geq 2$, $q \geq 1$ and $A > 0$ (for $p \in [1, 2)$ the upper bound includes a log factor).

Superefficiency of Block James-Stein estimator

Block James-Stein estimator δ^{BJS} is superefficient at every fixed point.

Theorem 4. *(Theorem 4 of Cai,2008). At any fixed point* $\theta \in b^r_{pq}(A)$ *with* $p \geq 2$ *and* $q < \infty$*, Block James-Stein estimator* δ^{BJS} *is superefficient, that is,*

$$\limsup_{n\to\infty} n^{2r/(1+2r)}\mathbb{E}||\delta - \theta^{BJS}||_2^2 = 0.$$

Summary: adaptivity and separability

As Cai (2008) shows (for $u = 2$),

- there exist adaptive estimators that achieve this rate without paying a penalty for adaptation, and they must be *nonseparable*, i.e. they must borrow information across its neighbours.
- For separable rules, the best possible rate has a log factor
- to achieve adaptability over $B^r_{p,q}(A)$, a nonseparable estimator should borrow information from at least $C[\log n]^{2r/(1+2r)}$ wavelet coefficients.

Conditions on adaptive estimator to be globally and locally minimax optimal

Non-adaptive estimators: Cai, Low and Zhao (2008) showed that it is not possible for a globally minimax optimal estimator to achieve the pointwise minimax optimal rate without paying a penalty $[\log n]^{u(r-1/p)/(2(r-1/p)+1)}$.

However, it is possible for an adaptive estimator to achieve the global and local adaptive minimax rates:

e.g. block James-Stein estimator (Cai,1999) achieves the optimal global rate with $u = 2$ over $B^r_{pq}(A)$ for all $r \in (0, s)$, $p \geq 2$, $q \geq 1$ and $A > 0$

and the optimal adaptive local rate (with $u = 2$) over $B^r_{\infty,\infty}$.

Empirical Bayes estimator

Johnstone & Silverman (2005) suggested an empirical Bayes estimator that achieves global minimax rate of convergence for $u \in (0, 2]$ over Besov spaces $B^r_{p,q}(A)$.

Scaling coefficients: $\hat{\theta}_k = y_k$ (observed values).

Wavelet coefficients.

Data: $y_{jk} \mid \theta_{jk} \sim \mathcal{N}(\theta_{jk}, \sigma^2/n)$ - independent, (σ^2 is known)

Prior: $\theta_{jk} \mid \pi_j \sim (1 - \pi_j)\delta_0 + \pi_j h$ - independent.

- The estimate $\hat{\theta}_{jk}$ of θ_{jk} is the **median of the posterior distribution** of θ_{jk}, with **plugged in** $\hat{\pi}_j$. It is a thresholding estimator with threshold $\hat{t}_j = t_j(\hat{\pi}_j)$.

  (It is a Bayes estimator for the dominating loss on θ for $u = 1$)

- Estimate π_j using maximum marginal likelihood approach.

- The corresponding estimator $\hat{f}_{EB}(t)$ has wavelet coefficients $\hat{\theta}_{jk}$.

Maximum marginal likelihood estimator for π_j

Marginal distribution of y_{jk} (given π_j) is

$$y_{jk} \mid \pi_j \sim (1-\pi_j)\varphi_{\sigma/\sqrt{n}} + \pi_j h^*, \quad k \in K_j.$$

where $h^*(x) = (h \star \varphi_{\sigma/\sqrt{n}})(x)$, and hence the maximum marginal likelihood estimator (MMLE) $\hat{\pi}_j$ satisfies

$$\hat{\pi}_j = \arg \max_{\pi_j \in [p_j,1]} \ell(\pi_j) = \arg \max_{\pi_j \in [p_j,1]} \sum_{k \in K_j} \log \left[(1-\pi_j)\varphi_{\sigma/\sqrt{n}}(y_{jk} + \pi_j h^*(y_{jk}) \right].$$

p_j is chosen in such a way that $t_j(p_j) \leq \sigma\sqrt{2\log(2^j)/n}$ for all $L \leq j < J$.

It can be viewed as an approximate Bayes estimator

Let $\pi_j \sim U[p_j, 1]$. Then the posterior distribution of π_j is

$$p(\pi_j \mid \mathbf{y}) = \frac{\prod_k p(y_{jk} \mid \pi_j)}{(1-p_j)p(y_{jk})} = \frac{1}{(1-p_j)p(y_{jk})} \exp\{\ell(\pi_j)\},$$

and achieves the maximum at $\hat{\pi}_j$.

The posterior distribution of θ_{jk} is

$$\begin{aligned} p(\theta_{jk} \mid \mathbf{y}) &= \int_{p_j}^{1} p(\theta_{jk} \mid \mathbf{y}, \pi_j) p(\pi_j \mid \mathbf{y}) d\pi_j \\ &\approx p(\theta_{jk} \mid \mathbf{y}, \hat{\pi}_j) \end{aligned}$$

using Laplace approximation.

Assumptions on the prior distribution

1. Density h is symmetric and unimodal
2. $\sup_{u>0} \left| \frac{d}{du} \log h(u) \right| < \infty$
3. $u^2 h(u)$ is bounded for all u
4. For some $\kappa \in [1, 2]$ and sufficiently large u, $\exists C_1, C_2 \in (0, \infty)$ such that

$$C_1 < \frac{1 - H(u)}{u^{\kappa - 1} h(u)} < C_2,$$

   where $H(x) = \int_{-\infty}^{x} h(u) du$.

Under these assumptions and Gaussian error, the posterior median $\hat{\theta}_{jk}$ is

- a thresholding rule (i.e. $\exists \hat{t}_j > 0$ such that $\hat{\theta}_{jk} = 0$ for $|y_{jk}| \leq \hat{t}_j$),
- with a bounded shrinkage property,i.e. $|\hat{\theta}_{jk} - y_{jk}| \leq \hat{t}_j + b$.

Global rate, upper bound

Theorem (Theorem 2 of Johnstone & Silverman, 2005).

Assume that ϕ *and* ψ *have regularity* s, $0 < p \leq \infty$, $0 < u \leq 2$, $r < s$, *and that either* $r > 1/p$ *or* $r = p = 1$. *Suppose that*
$\mathcal{F} = \{f : \quad$ *its wavelet coefficients* $\quad \theta \in b^r_{p,\infty}(A)\}$.

Then, there is a constant c *such that*

$$\sup_{f \in \mathcal{F}} \mathbb{E}||\hat{f} - f||_u^u \leq c[A\Lambda_{r,p}(nA^2) + A^u n^{-u(r-\nu)}[\log n]^\alpha + n^{-u/2}[\log n]^4].$$

where $\nu = \max[(1/p - 1/u)_+, (1/p - 1)_+]$, $\alpha \in \{0, 1\}$.

In particular, the theorem implies that EB estimator achieves minimax global rate for $u \in [1, 2]$ and $B^r_{p,q}$ with $p, q \in [1, \infty]$, $r < s$ and either $r > 1/p$ or $r = p = 1$.

Is it locally optimal?

Local minimax rate

Theorem. *Let* $0 < u \leqslant 2$, $r > 1/p$, $1 \leqslant p, q, \leqslant \infty$, $\phi, \psi \in L^\infty[0,1]$,
$\mathcal{F} = \{f :$ *its wavelet coefficients* $\theta \in b^r_{p,\infty}(A)\}$.

Let $\hat{f}$ *be the EB estimator of J&S (2005), under the stated assumptions.*

Then, for $n \geqslant n_0$ *and any* $t_0 \in (0,1)$,

$$\begin{aligned}
\sup_{f \in \mathcal{F}} \mathbb{E}|\hat{f}(t_0) - f(t_0)|^u &\leqslant C_R^u \left(\frac{n}{\log n}\right)^{-\frac{u(r-1/p)}{2(r-1/p)+1}} \\
&+ C_H^u \left(\frac{n}{\log n}\right)^{-u/2} (\log n)^{u\nu} \\
&+ C_A^u n^{-u(r-1/p)} Q_n^{\max(u,1)},
\end{aligned}$$

where $Q_n = 1/(2^{\min(u,1)/p} - 1)$ *if* $p < \infty$ *and* $Q_n = \log n / \log 2$ *if* $p = \infty$,
and rate $\nu = \frac{4 - \min(p,2)}{2\max(u,1)} + \mathbb{I}(u = 2) \in [1, 5/2]$.

Hence, EB estimator of Johnstone & Silverman (2005) also achieves the best possible local rate of convergence for $u \in [1,2]$ over $B^r_{p,q}$ with $r > 1/p$.

Local minimax rate: constants

The constants are given by

$$
\begin{aligned}
C_R &= ||\psi||_\infty^u K^{u^*} 3^{2u^*-1}[\tilde{A}^u + \sigma^u 2^{u/2+4}], \\
C_H &= 3^{2u^*-1} C_K^{u^*} c_u 2^{\frac{u-p_*}{2}} C_{g1}^{-1}, \\
C_A &= 3^{2u^*-1} C_K^{u^*} \tilde{A}^u \\
\nu &= \frac{4-\min(p,2)}{2u^*} + \mathbb{I}(u=2) \in [1, 5/2].
\end{aligned}
$$

where $u^* = \max(u,1)$, $u_* = \min(u,1)$, $\tilde{A} = c(\phi,\psi,r,p)A$,

$$C_K = \max(K, K_{L-1})\,[\max(||\phi||_\infty, ||\psi||_\infty)]^{u_*}\,[\max(\sigma,\sigma_{L-1})]^{u_*},$$

and $K = \sup_{j,t_0} |\{k :\, t_0 \in supp(\psi_{jk})\}|$.

Next steps

1. Are the assumptions on the prior distribution h necessary?

   More generally, can we find the full set of Bayesian models among

   $$\{\theta_{jk} \sim \pi_j h + (1 - \pi_j)\delta_0, \quad \pi_j \sim p(\pi_j)\}$$

   that such that the resulting posterior median estimator is simultaneously globally and locally (adaptively) minimax over $B^r_{p,q}$?

2. Are assumptions of the method realistic in order for the estimator to be applied in practice?

   Almost, the only unrealistic assumption is known variance. In the frequentist approach, commonly used estimator of σ^2 is $MAD(d_{J-1,k})/0.6745$.

One can consider alternative Bayesian estimators, e.g. conditional posterior mean that is also a thresholding estimator. Johnstone and Silverman (2005) studied a similar estimator, thresholded posterior mean, with the threshold from the posterior median. They found that such estimator is globally minimax over $B^r_{p,q}$ for $p \in (1, 2]$.

Sufficiency of assumptions on prior distribution

1. Density h is symmetric, unimodal

   If not satisfied, then the posterior median is not antisymmetric, not increasing.

2. $u^2 h(u)$ is bounded for all u

3. $\sup_{u>0} \left| \frac{d}{du} \log h(u) \right| < \infty$.

   Implies bounded shrinkage property of the posterior median. If it is not satisfied (e.g. for the Gaussian prior $w_{jk} \sim N(0, \tau^2 n)$), we have $\hat{\theta}_{jk} = \frac{\tau^2}{\tau^2 + \sigma^2/n} y_{jk}$ and the amount of shrinkage $|\hat{\theta}_{jk} - y_{jk}| = \frac{\sigma^2}{n\tau^2 + \sigma^2} |y_{jk}|$ is unbounded for large $|y_{jk}|$ and fixed n.

   Johnstone & Silverman (2005) show that in this case the maximum risk is a constant.

Sufficiency of assumptions on prior distribution

4. For some $\kappa \in [1, 2]$ and sufficiently large u, $\exists C_1, C_2 \in (0, \infty)$ such that

$$C_1 < \frac{1 - H(u)}{u^{\kappa-1} h(u)} < C_2,$$

where $H(x) = \int_{-\infty}^{x} h(u) du$.

Case $\kappa = 1$ corresponds to distributions with exponential tails, $\kappa = 2$ - to distribution with polynomial tails (Pareto - like).

Open problems

- Given a loss function, characterisation of prior distributions such that the corresponding posterior distribution is efficient.

  At least some robustness to misspecification of error distribution.

- Characterisation of prior distributions such that the posterior distribution is efficient in both “local” and “global” losses.

Summary

1. Lecture 1: Examples of classical and Bayesian estimators for nonparametric regression
2. Lecture 2: Application of minimax optimality to derive adaptive estimators with good performance, illustrated for local polynomial regression estimators
3. Lecture 3: Wavelet estimators for nonparametric regression, local and global minimax optimality
4. Lecture 4: Wavelet estimators that achieve local and global optimality simultaneously.

THANK YOU

for your attention.