Bayesian wavelet estimators in nonparametric regression

Natalia Bochkina

University of Edinburgh

Outline

Lecture 1. Classical and Bayesian approaches to estimation in nonparametric regression

1. Classical estimators
   - Kernel estimators
   - Orthogonal series estimators
   - Other estimators (local polynomials, spline estimators etc)
2. Bayesian approach
   - Prior on coefficients in an orthogonal basis
   - Gaussian process priors
   - Other prior distributions

Lecture 2. Classical minimax consistency and concentration of posterior measures

1. Decision-theoretic approach to classical consistency and concentration of posterior measures
2. Classical consistency
   - Bayes and minimax estimators
   - Speed of convergence
   - Adaptivity
   - Lower bounds
3. Concentration of posterior measures

Lecture 3. **Wavelet estimators in nonparametric regression**

1. Thresholding estimators (universal, SURE, Block thresholding; different types of thresholding functions).
2. Different choices of prior distributions
3. Empirical Bayes estimators (posterior mean and median, Bayes factor estimator)
4. Optimal non-adaptive wavelet estimators
5. Optimal adaptive wavelet estimators

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

Lecture 1. Classical and Bayesian approaches to estimation in nonparametric regression

1. Classical estimators
    - Kernel estimators
    - Orthogonal series estimators
    - Other estimators
2. Bayesian approach
    - Prior on coefficients in an orthogonal basis
    - Gaussian process priors
    - Other prior distributions

Lecture 1. Main references

- A. Tsybakov (2009) Introduction to nonparametric estimation. Springer.
- J. Ramsay and B. Silverman (2002) Functional data analysis. Springer
- B. Vidakovic (1999) Statiatical modeling via wavelets. Wiley.
- K. Rasmussen & C. Williams (2006) Gaussian processes for machine learning. MIT Press.

Examples of nonparametric models and problems

- Estimation of a probability density

  Let $X_1, \ldots, X_n \sim F$ iid, distribution F is absolutely continuous with respect to the Lebesgue measure μ on $\mathbb{R}$.

  Aim: estimate the unknown density $p(x) = \frac{dF}{d\mu}$.

- Nonparametric regression

  Assume pairs of random variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ are such that

$$Y_i = f(X_i) + \varepsilon_i, \quad X_i \in [0, 1],$$

  where $\mathbb{E}(\varepsilon_i) = 0$ for all i. We can write $f(x) = \mathbb{E}(Y_i \mid X_i = x)$.

  Unknown function $f : [0, 1] \to \mathbb{R}$ is called the *regression function*.

  The problem of nonparametric regression is to estimate unknown function f.

We focus on the nonparametric regression problem, large sample properties.

Examples of nonparametric models and problems (cont)

- White noise model

  This is an idealised model that provides an approximation to the nonparametric regression model. Consider the following stochastic differential equation:

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}}dW(t), \quad t \in [0,1],$$

  where W is a standard Wiener process on $[0,1]$, the function f is an unknown function on $[0,1]$, and n is an integer. It is assumed that a sample path $\{Y(t), 0 \leq t \leq 1\}$ of the process Y is observed.

The statistical problem is to estimate the unknown function f.

First introduced in the context of nonparametric estimation by Ibragimov and Hasminskii (1977, 1981)

Formally asymptotic equivalence was proved by Brown and Low (1996).

An extension to the multivariate case and random design regression was obtained by Reiss (2008).

Parametric vs nonparametric estimation

1. Parametric estimation

If we know a priori, that unknown f (regression function or density function) belongs to a parametric family $\{g(x,\theta):\ \theta\in\Theta\}$, where $g(\cdot,\cdot)$ is a given function, and $\Theta\subset\mathbb{R}^k$ (k is fixed, independent of n), then estimation of f is equivalent to estimation of the finite-dimensional parameter θ.

Examples: 1. Density p is normal $\mathcal{N}(a,\sigma^2)$, unknown parameter $\theta=(a,\sigma^2)\in\Theta=\mathbb{R}\times\mathbb{R}_+$.

2. Regression function $f(x)$ is linear: $f(x)=ax+b$, $\theta=(a,b)\in\Theta=\mathbb{R}^2$.

If such a prior information about f is not available we deal with a nonparametric problem.

Parametric vs nonparametric estimation

2. Nonparametric estimation

An ill-posed problem, hence usually additional prior assumptions on f are used.

Direct assumption: f belongs to some "massive" class $\mathcal{F}$ of functions. For example, $\mathcal{F}$ can be the set of all the continuous functions on $\mathbb{R}$ or the set of all differentiable functions on $\mathbb{R}$.

Tuning parameters of the estimators considered below are chosen to achieve best performance in the specified class of functions.

Indirect assumptions are also used, e.g. via penalisation or prior distribution on f in Bayesian approach.

Nonparametric regression estimators

1. Kernel estimators.

Density estimation: $X_1, \ldots, X_n$ - iid random variables with (unknown) density $p(x)$ wrt Lebesgue measure on $\mathbb{R}$.

The corresponding distribution function is $F(x) = \int_{-\infty}^{x} p(t)dt$.

The empirical distribution function

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x),$$

where $I(A)$ denotes the indicator function of set A. By the strong law of large numbers, we have

$$\hat{F}_n(x) \to F(x), \quad \forall x \in \mathbb{R},$$

almost surely as $n \to \infty$. Therefore, $\hat{F}_n(x)$ is a consistent estimator of $F(x)$ for every $x \in \mathbb{R}$. How can we estimate the density p?

Kernel density estimators (cont)

One of the first intuitive solutions is based on the following argument. For sufficiently small $h > 0$ we can write an approximation

$$p(x) = F'(x) \approx \frac{F(x+h) - F(x-h)}{2h}.$$

Replacing F by $\hat{F}_n$, we define

$$\hat{p}_n^R(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h}$$

which is called *Rosenblatt estimator*. It can be rewritten in the form

$$\hat{p}_n^R(x) = \frac{1}{2nh} \sum_{i=1}^{n} I(x-h < X_i \leq x+h) = \frac{1}{nh} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

where $K_0(x) = \frac{1}{2} I(-1 < x \leq 1)$.

Kernel density estimators (cont)

A simple generalisation of the Rosenblatt estimator is given by

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K : \mathbb{R} \to \mathbb{R}$ is an integrable function satisfying $\int K(u)du = 1$. Such a function K is called a *kernel* and the parameter h is called a *bandwidth* of the estimator $\hat{p}_n(x)$. The function $\hat{p}_n(x)$ is called the kernel density estimator or the *Parzen-Rosenblatt estimator*.

Further reading: B. Silverman (1986) Density estimation for statistics and data analysis. Wiley.

Tuning parameters: bandwidth h and kernel K.

Kernel estimators for regression function

Nonparametric regression model:

$$Y_i = f(X_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where (X_i, Y_i) are iid pairs, $\mathbb{E}|Y_i| < \infty$, $f(x) = \mathbb{E}(Y_i \mid X_i = x)$ - regression function to be estimated.

Given a kernel K and a bandwidth h, one can construct various kernel estimators for nonparametric regression similar to those for density estimation. The most celebrated one is the Nadaraya - Watson estimator.

Motivation for Nadaraya - Watson estimator

Suppose (X, Y) has density $p(x, y)$ with respect to the Lebesgue measure and $p(x) = \int p(x, y)dy > 0$. Then

$$f(x) = E(Y|X = x) = \frac{\int yp(y \mid x)dy}{p(x)} = \frac{\int yp(x, y)dy}{p(x)}.$$

If we replace here $p(x, y)$ by its kernel estimator $\hat{p}_n(x, y)$:

$$\hat{p}_n(x, y) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right),$$

and use the kernel estimator $\hat{p}_n(x)$ instead of $p(x)$, if kernel K is of order 1, we obtain Nadaraya-Watson estimator:

$$\hat{f}_n^{NW}(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}, \quad \text{if} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0,$$

and $\hat{f}_n^{NW}(x) = 0$ otherwise.

Nadaraya-Watson estimator is linear

The Nadaraya-Watson estimator can be represented as a weighted sum of Y_i:

$$\hat{f}_n^{NW}(x) = \sum_{i=1}^{n} Y_i W_{ni}^{NW}(x),$$

where the weights are given by

$$W_{ni}^{NW}(x) = \frac{K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} I\left(\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \neq 0\right).$$

Definition 1. *An estimator $\hat{f}_n(x)$ of $f(x)$ is called a linear nonparametric regression estimator if it can be written in the form*

$$\hat{f}_n(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

where the weights $W_{ni}(x) = W_{ni}(x, X_1, ..., X_n)$ depend only on n, i, x and the values $X_1, \ldots, X_n$.

Typically, $\sum_{i=1}^{n} W_{ni}(x) = 1$ for all x (or for almost all x wrt Lebesgue measure).

Nadaraya - Watson estimator (cont)

If density $p(x)$ of X_i is known, we can use it instead of $\hat{p}_n(x)$, then we obtain a different kernel estimator:

$$\hat{f}_n^{NW}(x) = \frac{1}{nhp(x)} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)$$

and, in case of uniform design ($X_i \sim U[0,1]$),

$$\hat{f}_n^{NW}(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right).$$

This estimator is also applicable for the regular fixed design $x_i = i/n$.

Other kernels

- $K(u) = (1 - |u|)I(|u| \leq 1)$ - triangular kernel
- $K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$ - parabolic, or Epanechnikov kernel
- $K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ - Gaussian kernel
- $K(u) = \frac{1}{2} e^{-|u|/\sqrt{2}} \sin(|u|/\sqrt{2} + \pi/4)$ - Silverman kernel.

Local polynomial estimators

If the kernel K takes only nonnegative values, the Nadaraya-Watson estimator $\hat{f}_n^{NW}$ satisfies

$$\hat{f}_n^{NW}(x) = \arg\min_{\theta \in \mathbb{R}} \left\{ \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h} \right) \right\}$$

Thus $\hat{f}_n^{NW}$ is obtained by a **local constant least squares approximation** of the outputs Y_i.

Local polynomial least squares approximation: replace constant θ by a polynomial of given degree k. If $\exists f^{(k)}$, then for z sufficiently close to x we may write

$$f(z) \approx f(x) + f'(x)(z-x) + \ldots + \frac{f^{(k)}(x)}{k!}(z-x)^k = \theta^T(x) U\left(\frac{z-x}{h} \right),$$

where

$$U(u) = (1, u, u^2/2!, \ldots, u^k/k!)^T, \quad \theta(x) = (f(x), f'(x)h, f''(x)h^2, \ldots, f^{(k)}(x)h^k)^T.$$

Local polynomial estimators

Definition 2. *Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel, $h > 0$ be a bandwidth, and $k > 0$ be an integer. A vector $\theta(x) \in \mathbb{R}^{k+1}$ defined by*

$$\arg \min_{\theta \in \mathbb{R}^{k+1}} \left\{ \sum_{i=1}^{n} \left[Y_i - \theta^T(x) U \left(\frac{z - x}{h} \right) \right]^2 K \left(\frac{X_i - x}{h} \right) \right\}$$

is called a local polynomial estimator of order k of $f(x)$. The statistic

$$\hat{f}_n(x) = U^T(0) \widehat{\theta}_n(x)$$

is called a local polynomial estimator of order k.

2. Projection estimators (orthogonal series estimators)

Nonparametric regression model:

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $\mathbb{E}\epsilon_i = 0$, $\mathbb{E}\epsilon_i^2 < \infty$.

Assume $x_i = i/n$, $f \in L^2[0,1]$.

Take some orthonormal basis $\{\varphi_k(x)\}_{k=0}^{\infty}$ of $L^2[0,1]$. Then, for any $f \in L^2[0,1]$, $\exists \{\theta_k\}_{k=0}^{\infty}$:

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x),$$

and $\theta_k = \int_0^1 f(x)\varphi_k(x)dx$.

Projection estimation of f is based on a simple idea: approximate f by its projection $\sum_{k=0}^{N} \theta_k \varphi_k(x)$ on the linear span of the first $N+1$ functions of the basis, and replace θ_k by their estimators.

Projection estimators

If X_i are scattered over $[0,1]$ in a sufficiently uniform way, which happens, e.g., in the case $X_i = i/n$, the coefficients θ_k are well approximated by the sums $\frac{1}{n}\sum_{i=1}^{n} f(X_i)\varphi_k(X_i)$.

Replacing in these sums the unknown quantities $f(X_i)$ by the observations Y_i we obtain the following estimators of θ_k :

$$\widehat{\theta}_k = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_k(X_i).$$

Definition 3. *Let $N \geq 1$ be an integer. The statistic*

$$\hat{f}_n^N(x) = \sum_{k=0}^{N} \widehat{\theta}_k \varphi_k(x)$$

is called a projection estimator (or an orthogonal series estimator) of the regression function f at the point x.

Choice of parameter N corresponds to choosing smoothness of f.

Projection estimators (cont)

Note that $\hat{f}_n^N(x)$ is a linear estimator, since we may write it in the form

$$\hat{f}_n^N(x) = \sum_{i=1}^{n} Y_i W_{ni}(x)$$

with

$$W_{ni}(x) = \frac{1}{n} \sum_{k=0}^{N} \varphi_k(x)\varphi_k(X_i)$$

Examples:

1. Fourier basis: $\varphi_{2k}(x) = 1$, $\varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx)$, $\varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx)$, $k = 1, 2, \ldots$, $x \in [0, 1]$ (Tsybakov, 2009).
2. A wavelet basis (Vidakovic, 1999)
3. An orthogonal polynomial basis: $\varphi_k(x) = (x - a)^k$, $k \geq 0$ (more commonly used in the context of density estimation)

Generalisation to arbitrary X_is

Define vectors $\theta = (\theta_0, \ldots, \theta_N)^T$ and $\varphi(x) = (\varphi_0(x), \ldots, \varphi_N(x))^T$.

The least squares estimator $\widehat{\theta}^{LS}$ of the vector θ is defined as follows:

$$\widehat{\theta}^{LS} = \arg\min_{\theta \in \mathbb{R}^N} \sum_{i=1}^{n} (Y_i - \theta^T \varphi(X_i))^2.$$

If the matrix

$$B = \sum_{i=1}^{n} \varphi(X_i)\varphi^T(X_i)$$

is invertible, we can write

$$\widehat{\theta}^{LS} = B^{-1} \sum_{i=1}^{n} Y_i \varphi(X_i).$$

Then the nonparametric least squares estimator of $f(x)$ is given by

$$\hat{f}_{n,N}^{LS}(x) = \varphi^T(x)\widehat{\theta}^{LS}.$$

Wavelet basis

Wavelet basis with periodic boundary correction on $[0, 1]$ is

$$\{\phi_{Lk}, k = 0, \ldots, 2^L - 1;\ \psi_{jk},\ j = L, L+1, \ldots, k = 0, \ldots, 2^j - 1\},$$

where $\phi_{jk}(x) = 2^{j/2}\phi(2^j x - k)$, $\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k)$,

$\phi(x)$ is a **scaling function**, $\psi(x)$ is a **wavelet function** such that

$$\int \phi(x)dx = 1, \quad \int \psi(x)dx = 0.$$

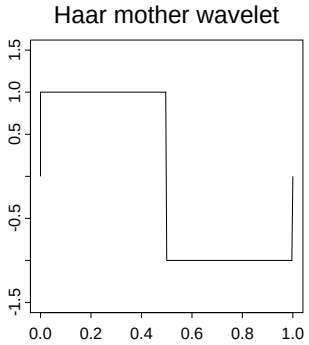
Then, any $f \in L^2[0, 1]$ can be decomposed in the **wavelet basis:**

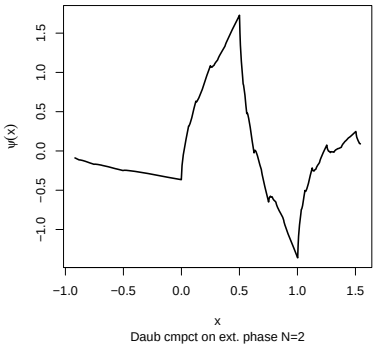
$$f(x) = \sum_{k=0}^{2^L-1} \theta_k \phi_{Lk}(x) + \sum_{j=L}^{\infty} \sum_{k=0}^{2^j-1} \theta_{jk}\psi_{jk}(x),$$

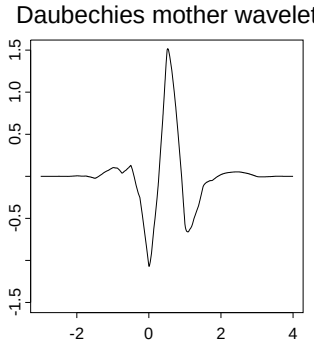
and $\theta = \{\theta_k, \theta_{jk}\}$ is a set of wavelet coefficients. [Meyer, 1990]

Wavelets (ϕ, ψ) are said to have regularity s if they have s derivatives and ψ has s vanishing moments ($\int x^k \psi(x)dx = 0$ for integer $k \leq s$).

Examples of wavelet functions



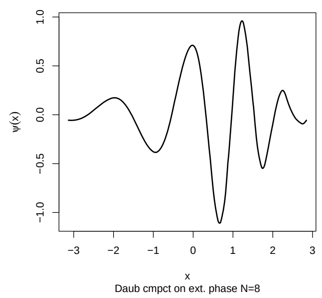


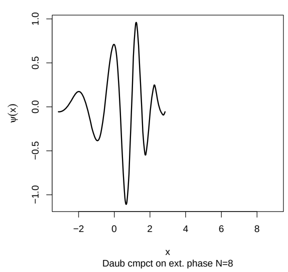


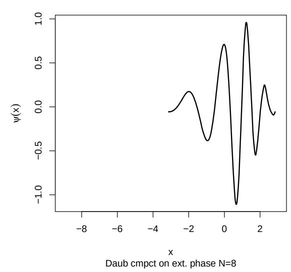
(a) Haar wavelet (b) Daubechies wavelet, $s = 4$ (c) Daubechies wavelet, $s = 2$

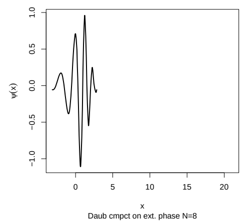
Localisation in time and frequency domains - sparse wavelet representation of most functions.

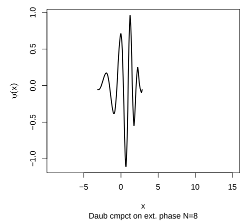
Daubechies wavelet transform, $s = 8$

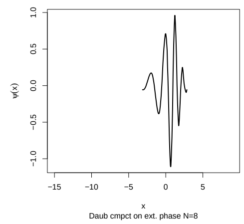


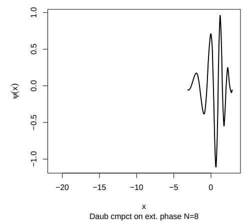












Discrete wavelet transform (DWT)

Applying discretised wavelet transform to data yields

$$\begin{aligned} d_{jk} &= w_{jk} + \varepsilon_{jk}, \quad L \leq j \leq J-1,\ k = 0, \ldots, 2^j - 1, \\ c_{Lk} &= u_{Lk} + \varepsilon_k, \quad k = 0, \ldots, 2^L - 1, \end{aligned}$$

where d_{jk} and c_{Lk} are discrete wavelet and scaling coefficients of observations (y_i), and ε_{jk} and ε_k are coefficients of the discrete wavelet transform of noise (ϵ_i). If $\epsilon_i \sim N(0, \sigma^2)$ independent, then $\varepsilon_{jk} \sim N(0, \sigma^2)$ independent.

Connection to θ_{jk}:

$$\theta_{jk} = \int_0^1 f(x)\psi_{jk}(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} \psi_{jk}(i/n)f(i/n) = \frac{1}{\sqrt{n}}(W f_n)_{(jk)} = \frac{w_{jk}}{\sqrt{n}} =: \tilde{\theta}_{jk},$$

where W is orthonormal $n \times n$ matrix, $f_n = (f(1/n), \ldots, f(1))$.

Also, for $y_{jk} = d_{jk}/\sqrt{n}$ and $y_k = c_{L,k}/\sqrt{n}$, and for Gaussian noise,

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

Smoothness

Fourier series - basis of Sobolev spaces $W_p^r \cap L^2$, $p \in [1, \infty]$, $r > 0$:

$$f \in W_p^r \quad \Leftrightarrow \quad \sum_{k=1}^{\infty} |a_k^r \theta_k|^p < \infty,$$

where $a_k = k$ for even k and $a_k = k - 1$ for odd k.

Wavelet series - basis of Besov spaces $B_{p,q}^r \cap L^2$), $p, q \in [1, \infty]$, $r > 0$:

$$f \in B_{p,q}^r \quad \Leftrightarrow \quad \left[\sum_{k=0}^{2^L - 1} |\theta_k|^p \right]^{1/p} + \left[\sum_{j=L}^{\infty} 2^{jq(r+1/2-1/p)} \left(\sum_{k=0}^{2^j - 1} |\theta_{jk}|^p \right)^{p/q} \right]^{1/q} < \infty$$

provided regularity s of wavelet transform: $s > r > 0$ (Donoho and Johnstone, 1998, Theorem 2).

Embeddings: $B_{2,2}^r = W_2^r$.

Regularisation

Penalised least squares estimator of f:

$$\hat{f}_n^{\text{pen}} = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} (Y_i - f(x_i))^2 + \lambda \text{pen}(f)$$

where pen(f) is a penalty function, $\lambda > 0$ is regularisation parameter.

Example: pen$(f) = \int [f''(x)]^2 dx$, leads to a cubic spline estimator (Silverman, 1985).

(see Green and Silverman, 1994, for more details).

Regularisation

Penalisation can be done on the coefficients of f in an orthonormal basis:

$$\hat{\theta}_n^{\text{pen}} = \arg \min_{\theta \in \mathbb{R}^{N+1}} \sum_{k=0}^{N} (y_k - \theta_k)^2 + \lambda \text{pen}(\theta)$$

Examples: 1. $\text{pen}(\theta) = ||\theta||_2^2$: $\widehat{\theta}_k = \frac{1}{1+\lambda} y_k$ - Tikhonov regularisation, ridge regression.

2. $\text{pen}(\theta) = ||\theta||_1$: for large enough λ, $\widehat{\theta}$ is sparse, lasso regression (Tibshirani, 1996).

Estimator $\hat{f}_n^{\text{pen}}$ ($\hat{\theta}_n^{\text{pen}}$) coincides with MAP (maximum a posteriori) Bayesian estimator.

Bayesian estimators

Likelihood:

$$Y_i = f(X_i) + \epsilon_i.$$

Common ways of specifying a prior distribution on a set of functions $\mathcal{F}$:

- On coefficients in some (orthonormal) basis, e.g. wavelet basis.
- Directly on $\mathcal{F}$, e.g. in terms of Gaussian processes

Inference is based on the posterior distribution $(f \mid Y)$:

$$p(f \min Y) = \frac{p(y \mid f)p(f)}{p(Y)}.$$

A point summary of the posterior distribution gives $\hat{f}$ (e.g. posterior mean, median, mode); can also obtain credibility bands for $\hat{f}$.

Bayesian projection estimators

Decomposition in some orthonormal basis:

$$f(x) = \sum_{k=0}^{\infty} \theta_k \varphi_k(x).$$

Likelihood under the (continuous time) white noise model:

$$Y_k \sim \mathcal{N}(\theta_k, \sigma^2/n) \quad \text{independent}$$

Under the nonparametric regression model: $y_k \sim \mathcal{N}(\tilde{\theta}_k, \sigma^2/n)$, independent.

Prior on coefficients θ:

$$\theta_k \sim p_k(\cdot), \quad k = 0, \ldots, N,$$

and $\mathbb{P}(\theta_k = 0) = 1$ for $k > N$.

Prior distributions π_k can be determined by a priori smoothness assumption.

Inference is based on the posterior distribution $\theta \mid y$: $\widehat{\theta}_k$ can be posterior mean, median, mode etc; variability of θ_k.

Example: posterior mode (MAP) estimator

Suppose we have Gaussian likelihood: $y_k \sim \mathcal{N}(\theta_k, \sigma^2/n)$, and prior densities $\theta_k \sim p_k(\cdot)$, $k = 0, \ldots, N$.

The corresponding posterior density of θ is

$$f(\theta \mid y) \propto \exp\left\{\sum_{k=0}^{N}[-\frac{n}{2\sigma^2}(y_k - \theta_k)^2 + \log p_k(\theta_k)]\right\}.$$

Posterior mode (MAP) estimator:

$$\hat{\theta}_n^{MAP} = \arg\max_{\theta \in \mathbb{R}^{N+1}} f(\theta \mid y) = \arg\min_{\theta \in \mathbb{R}^{N+1}} \sum_{k=0}^{N}(y_k - \theta_k)^2 + \lambda_n \text{pen}(\theta)],$$

where $\text{pen}(\theta) = -\sum_{k=0}^{N} \log p_k(\theta_k)$.

For example, for a Gaussian prior $\theta_k \sim \mathcal{N}(0, \tau^2)$ iid, $\text{pen}(\theta) = ||\theta||_2^2/2\tau^2$ - corresponds to ridge regression estimator, and for a double exponential prior $p_k(\theta_k) = \frac{\tau}{2}e^{-\tau|\theta_k|}$ iid, $\text{pen}(\theta) = \tau||\theta||_1$ - corresponds to lasso regression.

Choice of prior distribution for Bayesian wavelet estimators

Wavelet decomposition:

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

Wavelet representation of most functions is sparse, motivating the following prior distribution for wavelet coefficients:

$$\theta_{jk} \sim (1 - \pi_j)\delta_0(\cdot) + \pi_j h_j(\cdot),$$

where $h_j(\cdot)$ is the prior density function of non-zero wavelet coefficients, and $\pi_j = \mathbb{P}(\theta_{jk} \neq 0)$.

Scaling coefficients: $\theta_k \sim 1$ - noninformative prior.

Prior distribution of wavelet coefficients

h - normal: Clyde and George (1998), Abramovich, Sapatinas and Silverman (1998), etc.

h - double exponential: $h(x) = \frac{1}{2}e^{-|x|}$ - by Vidakovic (1998), Clyde and George (1998), Johnstone and Silverman (2005).

h - t distribution: Bochkina and Sapatinas (2005), Johnstone and Silverman (2005).

What is corresponding a priori regularity of f?

A priori regularity

Studied by Abramovich et al. (1998) for normal h and $\pi_j = \min(1, c_\pi 2^{-\beta j})$, $\tau_j = c_\tau 2^{-\alpha j}$, $\alpha, \beta \geqslant 0$, $c_\tau, c_\pi > 0$.

Generalised to arbitrary h, π_j and τ_j by Bochkina (2002)

[PhD thesis, University of Bristol]

Example: $\tau_j = c_\tau 2^{-\alpha j}$, $\pi_j = \min(1, c_\pi 2^{-\beta j})$.

Expected number of non-zero wavelet coefficients is $\mathbb{E}N = \sum_{j=j_0}^{\infty} 2^j \pi_j$.

Can specify π_j in such a way that:

$\mathbb{E}N = \infty$: $\pi_j = \min(1, C_\pi 2^{-\beta j})$ with $\beta \leq 1$;

$\mathbb{E}N < \infty$: $\pi_j = \min(1, C_\pi 2^{-\beta j})$ with $\beta > 1$.

Consider case $\beta \in (0, 1]$.

Assumptions on distribution H

Suppose ξ has distribution H.

1. $0 \leqslant \beta < 1, 1 \leqslant p < \infty, 1 \leqslant q \leqslant \infty$: assume that $E|\xi|^p < \infty$. If $q < \infty$, we also assume that $\mathbb{E}|\xi|^q < \infty$.

2. $0 \leqslant \beta < 1, p = \infty, 1 \leqslant q \leqslant \infty$: assume that distribution of $|\xi|$ has tail of one of the following types:

   (a) $1 - H(x) + H(-x) = c_l x^{-l}[1 + o(1)]$ as $x \to +\infty$, $l > 0$, $c_l > 0$; if $q < \infty$, assume that $l > q$;

   (b) $1 - H(x) + H(-x) = c_m e^{-(\lambda x)^m}[1 + o(1)]$ as $x \to +\infty$, $m > 0$, $\lambda > 0$, $c_m > 0$.

3. $\beta = 1, 1 \leqslant p \leqslant \infty, 1 \leqslant q < \infty$: assume that $\mathbb{E}|\xi|^q < \infty$.

4. $\beta = 1, 1 \leqslant p \leqslant \infty, q = \infty$: assume that $\exists \epsilon > 0$ such that $\mathbb{E}[\log(|\xi|) I(|\xi| > \epsilon)] < \infty$.

A priori regularity

$$\delta_H = \begin{cases} \frac{1-\beta}{l}, & H \quad \text{has polynomial tail and } p = \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Theorem 1. *Suppose that ψ and ϕ are wavelet and scaling functions of regularity s, where $0 < r < s$. Consider function f and its wavelet transform under assumption H.*

Then, for any fixed value of scaling coefficients θ_k, $f \in B^r_{p,q}$ almost surely if and only if

$$\text{either} \quad r + \frac{1}{2} - \frac{\alpha}{2} - \frac{\beta}{p} + \delta_H < 0,$$

$$\text{or} \quad r + \frac{1}{2} - \frac{\alpha}{2} - \frac{\beta}{p} = 0 \quad \text{and} \quad 0 \leqslant \beta < 1,\ p < \infty,\ q = \infty.$$

Nonparametric Bayesian estimators

Assume fixed design (i.e. $X_i = x_i$ are fixed):

$$Y_i = f(x_i) + \varepsilon_i, \quad x_i \in [0, 1],$$

with $\mathbb{E}(\varepsilon_i) = 0$ for all i.

Prior distribution: $f \sim \mathcal{G}$,

where $\mathcal{G}$ is a probability measure on a set of functions f.

Nonparametric Bayesian estimators: examples

1. $\mathcal{G} = \mathcal{GP}(m(x), k(x,y))$ - Gaussian process with mean function $m(x) = \mathbb{E}f(x)$ and covariance function $k(x,y) = Cov(f(x), f(y))$ - symmetric and positive definite.

2. Wavelet dictionary: Abramovich, Sapatinas, Silverman (2000), Bochkina (2002): model f as

$$f(x) = f_0(x) + f_w(x) = \sum_{i=1}^{M} \eta_{\lambda_i} \phi_{\lambda_i}(x) + \sum_{\lambda \in \Lambda} \omega_\lambda \varphi_\lambda(x),$$

where $\phi_\lambda(x) = a^{1/2}\phi(a(x-b))$, $\psi_\lambda(x) = a^{1/2}\psi(a(x-b))$
$\lambda = (a,b) \in [a_0, \infty) \times [0,1]$, $M < \infty$ and $\lambda_i < a_0$.

Take Λ - Poisson process on $\mathbb{R}_+ \times [0,1]$ with intensity $\mu(a,b) \propto a^{-\alpha}$, $\alpha > 0$, and $\omega_\lambda \mid \Lambda \sim H_\lambda(\cdot)$ iid.

For Gaussian H_λ, Abramovich et al. (2000) give necessary and sufficient conditions for $f \in B^r_{p,q}$ with probability 1, for more general H - in Bochkina (2002).

3. Levy adaptive regression kernels: $f(x) = \int g(x, \omega)\mathcal{L}(d\omega)$,

   where $\mathcal{L}(\omega)$ is a Levy random measure:

$$\mathcal{L}(A) = \sum_{k=0}^{N} \theta_k I_A(\omega_j)$$

   where $N \sim Pois(\mu)$, $(\beta_j, \omega_j) \sim \pi(d\beta, d\omega)$ iid (C. Tu, M.Clyde, R. Wolpert, 2007).

Nonparametric Bayesian estimators with Gaussian process prior

Definition 4. *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Assume that the observation errors are also Gaussian: $Y_i \sim \mathcal{N}(f(x_i), \sigma^2)$, or, in the matrix form,

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{f}, \sigma^2 I_n),$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$.

Often, in regression problems a priori $\mathbb{E}f(x) = m(x) = 0$.

Prior: $f \sim \mathcal{GP}(0, k(x, y))$.

Posterior distribution

Then, the posterior distribution of f at an arbitrary set of points $\mathbf{x}^* = (x_1^*, \ldots, x_m^*)^T \in (0,1)^m$, $\mathbf{f}^* = (f(x_1^*), \ldots, f(x_m^*))^T$ is

$$\mathbf{f}^* | \mathbf{Y}, \mathbf{x}, \mathbf{x}^* \sim \mathcal{N}_m(\mu, \Sigma)$$

where

$$\begin{aligned} \mu &= k(\mathbf{x}^*, \mathbf{x})[k(\mathbf{x}, \mathbf{x}) + \sigma^2 I_n]^{-1} \mathbf{Y}, \\ \Sigma &= k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*, \mathbf{x})[k(\mathbf{x}, \mathbf{x}) + \sigma^2 I_n]^{-1} k(\mathbf{x}^*, \mathbf{x}). \end{aligned}$$

If the posterior mean is used as a point estimator, we have, for any $x \in (0,1)$:

$$\hat{f}(x) = \mathbb{E}(f(x) | \mathbf{Y}, \mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(x_i, x),$$

where $\alpha = [k(\mathbf{x}, \mathbf{x}) + \sigma^2 I_n]^{-1} \mathbf{Y}$.

This estimator is *linear*, and is a particular case of kernel estimator.

In addition, have posterior credible bands.

Bayesian nonparametric estimators with Gaussian process prior

Smoothness

If we assume $f \in \mathcal{GP}(0, k(x, y))$, then $f \in \mathcal{H}_k$ - Reproducing Kernel Hilbert Space (RKHS) with kernel $k(x, y)$.

Hence, a priori regularity of a GP f is the regularity of the corresponding RKHS $\mathcal{H}$.

Orthogonal basis estimators with basis $\{\psi_i(x)\}$ are also (implicitly) assumed to belong to a RKHS with reproducing kernel $k(x, y) = \sum_{i=1}^{\infty} \psi_i(x)\psi_i(y)$.

Connection to splines:

if $k(x, y)$: $||f||^2_{\mathcal{H}} = \int [f''(x)]^2 dx$, the corresponding MAP estimator is a cubic spline.

The corresponding $k(x, y) = \frac{1}{2}(x - y)^2 \min(x, y) + \frac{1}{3}[\min(x, y)]^3$.

Regularity of Gaussian processes

1. Brownian motion: $k(x,y) = \frac{1}{2}[x + y - |x - y|]$.

$W(t) \in \mathbb{C}[0,1]$, $||W||^2_{\mathcal{H}} = W(0)^2 + ||W'||^2_2$.

2. Fractional Brownian motion: $k(x,y) = \frac{1}{2}[x^{2\alpha} + y^{2\alpha} - |x - y|^{2\alpha}]$,
$\alpha \in (0,1)$. α-smooth.

References:

- Q Wu, F Liang, S Mukherjee, RL Wolpert (2007) Characterizing the function space for Bayesian kernel models. Journal of Machine Learning.
- A. van der Vaart and H. Zanten (2008) Rates of contraction of posterior distributions based on Gaussian process priors. Annals of Statistics (36).

Next lecture

Frequentist behaviour of nonparametric estimators:

- Consistency of (point) estimators $\hat{f}_n$.
- Concentration of posterior measures.