Stat797 project: Nonparametric density estimation

Tobias Kuhlmann, Rui Zhang

12/12/2018

Data

For our project we simulate univariate data

$$\{y_i\}_{i=1}^{n} \quad i\epsilon\{1, ..., n\}$$

$y_i \sim iid$ and an unknown and known (different cases) smooth density f(x), i.e., $y_i \sim_{iid} f(x)$, where $x \in R$. This may not be just one dataset, but several.

Models

Univariate kernel density estimator

We use a univariate kernel density function following Wand and Jones (1995). A density function can be estimated by

$$\hat{f}(x; h) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\},$$

where K is a kernel function satisfying $\int K(x)dx = 1$ and h is the bandwidth.

MISE

$$MISE\{\hat{f}(\cdot; h)\} = E \int \{\hat{f}(x; h) - f(x)\}^2 dx$$

h_{MISE} is the minimiser of $MISE\{\hat{f}(\cdot; h)\}$ then

$$h_{MISE} \sim \big[\frac{R(K)}{\mu(K)^2 R(f'')n}\big]^{\frac{1}{5}} = C_1 n^{-\frac{1}{5}}$$

$$\text{inf MISE}_{h>0}\{\hat{f}(; h)\} \sim \frac{5}{4}\{\mu_2(K)^2 R(K)^4 R(f'')\}^{\frac{1}{5}} n^{-\frac{4}{5}} = C_2 n^{-\frac{4}{5}}$$

These expressions give the rate of convergence of the MISE-optimal bandwidth and the minimum MISE to zero as $n \to \infty$

Asymptotic MISE approximations can also be used to make comparisons of the kernel estimator to the histogram.

$$b_{MISE} \sim \{6/R(f')\}^{\frac{1}{3}} n^{-\frac{1}{3}}$$

$$\text{inf MISE}_{b>0}\{\hat{f}(\cdot; b)\} \sim \frac{1}{4}\{36R(f')\}^{\frac{1}{3}} n^{-\frac{2}{3}}$$

$$\text{MISE} = C_3 n^{-\frac{2}{3}}$$

$$\log(MISE) = -\frac{2}{3}\log(C_3 n)$$

Thus, the MISE of the histogram is asymptotically inferior to the kernel density estimator since its convergence rate is $O(n^{-\frac{2}{3}})$.

Univariate density estimation with logspline

Let B be a collection of feasible column vectors following Stone, Hansen, Kooperberg, and Truong (1997). If $\beta \epsilon B$, then

$$f(x;\beta) = exp(\beta_1 B_1(x) + \cdots + \beta_J B_J(x) - C(\beta)), L < x < U$$

where

$$C(\beta) = log(\int_L^U exp(\beta_1 B_1(x) + \cdots + \beta_J B_J(x))dx).$$

Then $f(y;\beta)$ is a positive density function on (L,U), and $\int_R f(x;\beta)dx = 1$.

As one of the penalized approaches, logspline uses a maximum likelihood approach.

Simulation experiment

Simulation

Monte Carlo experiment

Visualization

Goal and conclusion

After estimating both models on several sets of simulated data with different sample sizes, our goal is to study and compare the rates of convergence of the MISE as $n \to \infty$.