On the risk of estimates for block decreasing densities

Gérard Biau & Luc Devroye

Université Paris VI & McGill University

Summary

1. Introduction
2. Lower bounds
3. Minimax optimal estimates
4. Adaptive estimation

Introduction

Definition. *A density* $f = f(x_1, \ldots, x_d)$ *on* $[0, \infty)^d$ *is* **block decreasing** *if, for each* $j \in \{1, \ldots, d\}$*, it is a decreasing function of* x_j*, when all other components are held fixed.*

Examples. Burr, Pareto, beta conditional (Arnold, Castillo and Sarabia, 1999).

The problem.

$$\mathcal{R}_n(\mathcal{F}_B) = \inf_{f_n} \sup_{f \in \mathcal{F}_B} \mathbf{E} \int |f_n - f| \, .$$

References. Birgé (1987) studies the case $d = 1$ and shows that lower and upper bounds for the risk are proportional to $(S/n)^{1/3}$, where $S = \log(1 + B)$.

Lower bounds

Theorem. *There exist positive constants C_1, C_2 and C_3, functions of d, such that*

$$\mathcal{R}_n(\mathcal{F}_B) \geq \frac{1}{4\Big[1 + \big[1 + \big(C_1 S^d/n\big)^{\frac{1}{d+2}}\big]^{1/d}\Big]^d} \left(\frac{C_1 S^d}{n}\right)^{1/(d+2)}$$

for $C_2 \leq S \leq C_3\, n^{1/d}$.

Sketch of proof. Let ϵ be a positive real number and let $r \geq 1$ be an integer, both to be determined later. We partition the unit hypercube I_d into r^d cells

$$\mathcal{C}_{\mathbf{i}} = \prod_{j=1}^{d} [x_{i_j - 1}, x_{i_j}), \quad \mathbf{i} = (i_1, \ldots, i_d),$$

where $x_0 = 0$ and, for j in $\{1, \ldots, d\}$,

$$x_{i_j} = \frac{(1+\epsilon)^{i_j} - 1}{(1+\epsilon)^r - 1}, \quad i_j = 1, \ldots, r.$$

Birgé's multivariate histogram estimate

It is defined by

$$f_n = \sum_{\mathbf{i}} \frac{\mu_n(\mathcal{C}_{\mathbf{i}})}{\lambda(\mathcal{C}_{\mathbf{i}})} \mathbf{1}_{\mathcal{C}_{\mathbf{i}}} .$$

Theorem. *Birgé's multivariate histogram on I_d with*

$$r = \left\lceil \left(R^2 n S^2\right)^{1/(d+2)} \right\rceil, \quad R = \frac{2 + 2^{d-3}(d-1)}{\sqrt{2^d - 1}} ,$$

and

$$\epsilon = e^{S/r} - 1$$

satisfies

$$\sup_{f \in \mathcal{F}_B} \mathbf{E} \int |f_n - f| \le C_1 \left(\frac{R^d S^d}{n} \right)^{1/(d+2)} + C_2 \left(\frac{R^d S^d}{n} \right)^{2/(d+2)}$$

for all $n \ge C_3 S^d$, where C_1, C_2 and C_3 are positive functions of d.

A variable kernel estimate

It is defined by

$$f_n(x) = \frac{1}{n\tilde{\mathbf{h}}(x)} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\mathbf{h}(x)}\right),$$

where

$$K = \mathbf{1}_{[-1/2,1/2]^d} \quad \text{and} \quad \tilde{\mathbf{h}}(x) = h(x_1)\ldots h(x_d),$$

with

$$h(u) = \frac{\epsilon}{1+\frac{\epsilon}{2}}\left[u + \frac{1}{(1+\epsilon)^r - 1}\right].$$

Theorem. *The variable kernel estimate on I_d with*

$$r = \left\lceil \left(R^2 n S^2\right)^{1/(d+2)} \right\rceil, \qquad R = \frac{3 + 2^{d-3}(d-1)}{\sqrt{6^d}},$$

and

$$\epsilon = e^{S/r} - 1$$

satisfies

$$\sup_{f\in\mathcal{F}_B} \mathbf{E} \int |f_n - f| \leq C_1\left(\frac{R^d S^d}{n}\right)^{1/(d+2)} + C_2\left(\frac{R^d S^d}{n}\right)^{2/(d+2)}$$

for all $n \geq C_3 \max\left(S^d, S^{-2}\right)$, where C_1, C_2 and C_3 are positive functions of d.

Adaptive estimation

Fact. The estimate f_n depends upon the unknown parameter B, and more generally upon the parameter

$$\gamma(f) = f(0) \prod_{j=1}^{d} s_j,$$

with

$$\text{supp f} = [0, s_1] \times \ldots \times [0, s_d].$$

An automatic selection procedure for selecting the unknown parameters of a kernel estimate of the type

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} \frac{1}{h_j(x)} K\left(\frac{x^{(j)} - X_i^{(j)}}{h_j(x)}\right),$$

where

$$h_j(x) = \sum_{\ell=0}^{k} a(j,\ell) \left(x^{(j)}\right)^{\ell}, \quad a(j,\ell) \in \mathbb{R}.$$

► We want to select Θ in

$$\Omega = \left\{\theta = \big(a(j,\ell)\big), 1 \le j \le d, 0 \le \ell \le k\right\}.$$

Results

Theorem. *There exist two constants C_1 and C_2 (functions of d and k only) such that*

$$\mathbf{E}\int |f_{n,\Theta} - f| \le C_1 \inf_{\theta\in\Omega} \mathbf{E}\int |f_{n,\theta} - f| + C_2\sqrt{\frac{\log n}{n}}\,.$$

Corollary. *Let $\mathcal{F}$ denote the class of all block decreasing densities on $[0,\infty)^d$. The following adaptive result holds:*

$$\sup_{t>0}\ \sup_{\{f\in\mathcal{F}:\gamma(f)\le t\}} \frac{\mathbf{E}\int |f_{n,\Theta} - f|}{\left[\big(\log(t+1)\big)^d/n\right]^{1/(d+2)}} = \mathrm{O}(1).$$

The algorithm

1. Let $\Omega = \{\theta = \big(a(j,\ell)\big), 1 \le j \le d, 0 \le \ell \le k\}$.

2. Split the sample in two parts

$$X_1, \ldots, X_{n-m} \quad \text{and} \quad X_{n-m+1}, \ldots, X_n.$$

3. Consider the **Yatracos class** defined by

$$\mathcal{A} = \Big\{\{f_{n-m,\theta} > f_{n-m,\theta'}\}, \theta, \theta' \in \Omega, \theta \neq \theta'\Big\}.$$

4. Select Θ so as to minimize

$$\Theta = \arg\min \sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right| ,$$

where

$$\mu_m(A) = \frac{1}{m} \sum_{j=n-m+1}^{n} \mathbf{1}_{[X_j \in A]}.$$

Results II

Theorem (Devroye and Lugosi, 2001). *For $m \le n/2$, one has*

$$\mathbf{E}\int |f_{n,\Theta} - f| \le 5 \inf_{\theta\in\Omega} \mathbf{E}\int |f_{n,\theta} - f| \Big(1 + \frac{2m}{n-m} + 4\sqrt{\frac{m}{n}}\Big) + 8\mathbf{E}\Big\{\sup_{A\in\mathcal{A}} \Big|\int_A f - \mu_m(A)\Big|\Big\} + \frac{5}{n}\,.$$

Theorem (Vapnik and Chervonenkis, 1971). *We have*

$$\mathbf{E}\Big\{\sup_{A\in\mathcal{A}} \Big|\int_A f - \mu_m(A)\Big|\Big\} \le 2\sqrt{\frac{\log 2\,\mathcal{S}_{\mathcal{A}}(m)}{m}}\,.$$

Here, $\mathcal{S}_{\mathcal{A}}(m)$ is the **shatter coefficient** *for $\mathcal{A}$, i.e.,*

$$\mathcal{S}_{\mathcal{A}}(m) = \max_{(y_1,\ldots,y_m)\in\mathbb{R}^{dm}} \mathrm{Card}\big\{(y_1,\ldots,y_m)\cap A, A\in\mathcal{A}\big\}.$$

Lemma. *Set $l = n - m$. One has*

$$\mathcal{S}_{\mathcal{A}}(m) \le 2^{(k+1)^d+2+2(k+1)d} \times (d\ell)^{2(k+1)d} \times m^{2(k+1)d+(k+1)^d}\,.$$

Sketch of proof

- $x_1, \ldots, x_\ell \quad \rightarrow \quad$ The sample from $\mathbb{R}^{dl}$ used in the definition of $f_{\ell,\theta}$.
- $y_1, \ldots, y_m \quad \rightarrow \quad$ The test sample from $\mathbb{R}^{dm}$ to be employed.
- We begin by defining the vector

  $$V(j,i,\theta) \stackrel{\text{def}}{=} \left(K\left(\frac{y_j^{(1)} - x_i^{(1)}}{h_1(y_j)}\right), \ldots, K\left(\frac{y_j^{(d)} - x_i^{(d)}}{h_d(y_j)}\right) \right)$$

  and the $m \times \ell$ matrix $V(\theta)$ of vectors $V(j,i,\theta)$.
- There exists a partition of Ω^2 of size at most

  $$\left(2(2dm\ell)^{(k+1)d}\right)^2$$

  such that on any set of the partition, $\big(V(\theta), V(\theta')\big)$ is fixed.
- We are interested in the collection of indicators

  $$\left(\mathbf{1}_{f_{\ell,\theta}(y_1) > f_{\ell,\theta'}(y_1)}, \ldots, \mathbf{1}_{f_{\ell,\theta}(y_m) > f_{\ell,\theta'}(y_m)}\right).$$

- The shatter coefficient is bounded by

  $$W \times \left(2(2dm\ell)^{(k+1)d}\right)^2.$$

- To compute W, we fix all values of

$$V(j,i,\theta) = \left(K\left(\frac{y_j^{(1)} - x_i^{(1)}}{h_1(y_j)}\right), \ldots, K\left(\frac{y_j^{(d)} - x_i^{(d)}}{h_d(y_j)}\right)\right).$$

- Thus, $\left\{f_{\ell,\theta}(y) > f_{\ell,\theta'}(y)\right\}$ is a set defined by an inequality of the form

$$\prod_{s=1}^{d} h'_s(y) > c \prod_{s=1}^{d} h_s(y)\,.$$

- This is a **polynomial** inequality with each monomial being of the form

$$\left(y^{(1)}\right)^{p_1} \times \cdots \times \left(y^{(d)}\right)^{p_d}.$$

▶ The number of such monomials does not exceed $r = (k+1)^d$.

- Considered as a set in $\mathbb{R}^r$,

$$\prod_{s=1}^{d} h'_s(y) > c \prod_{s=1}^{d} h_s(y)$$

is just a homogeneous linear inequality of the form $a_1 w_1 + \cdots + a_r w_r > 0$.

- Therefore

$$W \le (m+1)^{(k+1)^d}\,.$$

References

[1] L. Birgé (1987a). Estimating a density under order restrictions: nonasymptotic minimax risk, *The Annals of Statistics*, **Vol. 15**, pp. 995–1012.

[2] L. Birgé (1987b). On the risk of histograms for estimating decreasing densities, *The Annals of Statistics*, **Vol. 15**, pp. 1013–1022.

[3] L. Devroye and G. Lugosi (2001). *Combinatorial Methods in Density Estimation*, Springer–Verlag, New York.

[4] http://www.ccr.jussieu.fr/lsta/biau.html