Slide 1

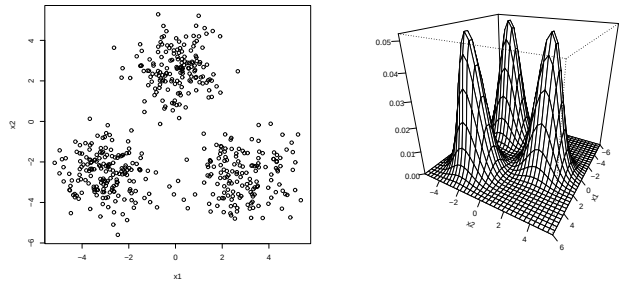
Density Estimation with Stagewise Optimization of the Empirical Risk

Jussi Klemelä, University of Mannheim

- Density estimation with iid vectors $X^1, \ldots, X^n \in \mathbf{R}^d$
- Minimizing the L_2 empirical risk in a stagewise manner
- Implementation with greedy histograms

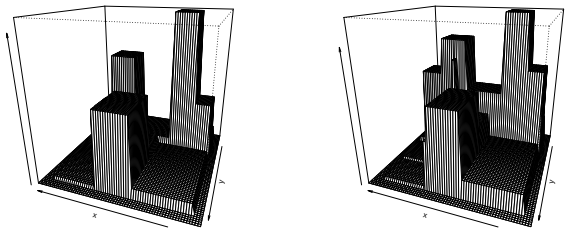
Slide 2

Data



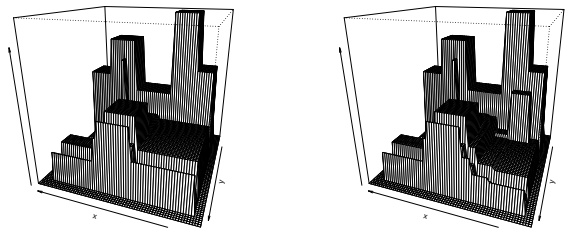
Slide 3

Stagewise minimization estimator



Slide 4

Stagewise minimization estimator II



Slide 7

Comments

1. The definition of the estimator is semi-algorithmic because the definition contains further minimization problems.
2. We may write the quantity to be minimized as

$$\gamma_n\left((1-\pi_k)\hat{f}_{k-1}+\pi_k\phi\right)$$
$$= \gamma_n\left((1-\pi_k)\hat{f}_{k-1}\right)+\gamma_n\left(\pi_k\phi\right)+2(1-\pi_k)\pi_k\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi.$$

   Thus we try to choose the new additive component $\pi_k\phi$ in such a way that the usual empirical risk $\gamma_n(\pi_k\phi)$ with an additional "penalization term" which involves the inner product $\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi$ is minimized.
3. The smoothing parameter of the estimate is the sequence $(\pi_k)_{k=1,\ldots,M-1}$.

Slide 8

The basic idea

We approximate the "ideal estimator"

$$\hat{f} = \text{argmin}_{g\in\text{co}(\mathcal{D})}\gamma_n(g)$$

with the stagewise procedure, with accuracy M^{-1}, where $\text{co}(\mathcal{D})$ is the convex hull of $\mathcal{D}$ (the collection of finite convex combinations of $\mathcal{D}$).

Slide 5

L_2 empirical risk

Define the empirical risk of a density estimator $\hat{f}:\mathbf{R}^d\to\mathbf{R}$ with

$$\gamma_n\left(\hat{f}\right)=\frac{1}{n}\sum_{i=1}^{n}\gamma\left(\hat{f},X^i\right)$$

where $\gamma(g,x)$ is the L_2 contrast function,

$$\gamma(g,x)=-2g(x)+\|g\|_2^2,\qquad g:\mathbf{R}^d\to\mathbf{R},\ \ x\in\mathbf{R}^d.$$

Note that

$$\begin{aligned}\|\hat{f}-f\|_2^2-\|f\|_2^2 &= -2\int_{\mathbf{R}^d}f\hat{f}+\|\hat{f}\|_2^2\\ &\approx -\frac{2}{n}\sum_{i=1}^{n}\hat{f}(X^i)+\|\hat{f}\|_2^2\\ &= \gamma_n\left(\hat{f}\right).\end{aligned}$$

Slide 6

Stagewise minimization estimator $\hat{f}_n:\mathbf{R}^d\to\mathbf{R}$

with dictionary $\mathcal{D}$ of functions $\mathbf{R}^d\to\mathbf{R}$,
with number of components $M\geq 1$,
with mixing coefficients $0<\pi_k<1$, $k=1,2,\ldots,M-1$,
with the approximation bound $\epsilon>0$:

1. Choose $\tilde{f}_0\in\mathcal{D}$ so that $\gamma_n(\tilde{f}_0)\leq\inf_{\phi\in\mathcal{D}}\gamma_n(\phi)+\epsilon$.
2. For $k=1,\ldots,M-1$, let

$$\tilde{f}_k=(1-\pi_k)\tilde{f}_{k-1}+\pi_k\tilde{\phi}$$

   where $\tilde{\phi}\in\mathcal{D}$ is chosen so that

$$\gamma_n(\tilde{f}_k)\leq\inf_{\phi\in\mathcal{D}}\gamma_n\left((1-\pi_k)\tilde{f}_{k-1}+\pi_k\phi\right)+\pi_k\epsilon.$$

3. Let $\hat{f}_n=\tilde{f}_{M-1}$.

Previous literature

1. Jones (1992): the minimization over the convex hull may be solved in a stagewise manner in the Hilbert space context
2. Barron (1993), Breiman (1993), Lee, Bartlett and Williamson (1996) refined the procedure in the L_2 regression estimation context.
3. A result of Maurey given in Pisier (1981) is referred as an origin of this algorithm.
4. Li and Barron (2000) considered density estimation with the log-likelihood empirical risk.
5. Zhang (2003) considers a general version of the algorithm.

A bound for the MISE

1. Let $\mathcal{F} = \bar{\text{co}}(\mathbb{G}) \cap \mathcal{F}_{den}$, where $\mathbb{G} \subset \{g : \mathbf{R}^d \to \mathbf{R}\}$, $\bar{\text{co}}(\mathbb{G})$ is the L_2 closure of the convex hull of $\mathbb{G}$, and $\mathcal{F}_{den}$ is the collection of bounded densities on $\mathbf{R}^d$;
2. let dictionary $\mathcal{D}$ be a finite δ-cover of $\mathbb{G}$, $\delta > 0$;
3. let the mixing coefficients be

$$\pi_k = \frac{2}{k+1}, \qquad k = 1, 2, \ldots, M-1.$$

Theorem 1 *Let $X^1, \ldots, X^n$ be i.i.d. observations from the distribution of density $f : \mathbf{R}^d \to \mathbf{R}$. For sufficiently large n, for $f \in \mathcal{F}$,*

$$E_f \left\| \hat{f}_n - f \right\|_2^2 \leq \delta^2 + C \left(\frac{\log_e(\#\mathcal{D})}{n} \right)^{1/2} + \frac{C}{M} + \epsilon + Cn^{-1/2}.$$

Basic application

Assume collection $\mathbb{G}$ has a δ-cover $\mathcal{D}_\delta$ of cardinality

$$\log_e(\#\mathcal{D}_\delta) \leq C \log_e \delta^{-1},$$

for a positive constant C. Then for a certain choice of the parameters of the estimator

$$\limsup_{n\to\infty} \left(\frac{n}{\log_e n} \right)^{1/2} \sup_{f \in \mathcal{F}} E_f \left\| \hat{f}_n - f \right\|_2^2 < \infty.$$

The classical nonparametric minimax rate is $n^{s/(2s+d)}$ where $s > 0$ is the smoothness index of the density class. We have that

$$(n/\log_e n)^{1/4} > n^{s/(2s+d)} \Leftrightarrow d > 2s.$$

Thus, in the high dimensional cases, when $d > 2s$, the rate $(n/\log_e n)^{1/4}$ is better than the classical rate.

Examples

- Let $\mathcal{F} = \bar{\text{co}}(\mathbb{G}) \cap \mathcal{F}_{den}$, where $\mathbb{G} = \{g(\cdot, \theta) : \theta \in \Theta\}$, $\Theta \subset \mathbf{R}^k$ and the parametrization is smooth. For example

$$g(x, \theta) = g(x, \mu, \sigma) = (2\pi)^{-d/2} \sigma^{-d} \exp\left\{ \frac{-\|x - \mu\|^2}{2\sigma^2} \right\}, \qquad x \in \mathbf{R}^d.$$

- Let $\mathcal{F} = \mathcal{G}(\mathbb{G}) \cap \mathcal{F}_{den}$, where

$$\mathcal{G}(\mathbb{G}) = \left\{ \int_{\Theta^{(2)}} g\left(\cdot, \theta^{(1)}, \theta^{(2)}\right) dQ\left(\theta^{(2)}\right) : \theta^{(1)} \in \Theta^{(1)}, Q \in \mathbb{Q}\left(\Theta^{(2)}\right) \right\}$$

(For example, the Gaussian mixture family)

- Let

$$\mathcal{F} = \mathcal{F}_{s,L} = \{g : \rho_s(g) \leq L\} \cap \mathcal{F}_{den},$$

where $\rho_s(f) = (2\pi)^d \int_{\mathbf{R}^d} \|\omega\|^s |F(f)(\omega)| \, d\omega$, $s \geq 0$. (L_1 Fourier class)

Slide 13

Proof of the theorem

Denote

$$f(x,\Lambda) = \sum_{\phi\in\mathcal{D}} \lambda_\phi \phi(x), \qquad x \in \mathbf{R}^d,$$

where $\Lambda = (\lambda_\phi)_{\phi\in\mathcal{D}} \in \mathcal{W}$, and $\mathcal{W}$ is the set of vectors of coefficients of convex combinations of $\mathcal{D}$.

1. We have for the estimator $\hat{f}_n$ that

$$\gamma_n\left(\hat{f}_n\right) \le \inf_{\Lambda\in\mathcal{W}} \gamma_n\left(f(\cdot,\Lambda)\right) + \frac{C}{M} + \epsilon,$$

2. When $\gamma_n\left(\hat{f}\right) \le \inf_{\Lambda\in\mathcal{W}} \gamma_n\left(f(\cdot,\Lambda)\right) + \varepsilon$, then

$$\left\|\hat{f} - f\right\|_2^2 \le \inf_{\Lambda\in\mathcal{W}} \|f(\cdot,\Lambda) - f\|_2^2 + \varepsilon + 4\sup_{\phi\in\mathcal{D}} |\nu_n(\phi)|,$$

where f is the true density, and $\nu_n(\phi)$ is the centered empirical

Slide 14

operator,

$$\nu_n(\phi) = \frac{1}{n}\sum_{i=1}^n \phi(X^i) - \int_{\mathbf{R}^d} \phi f, \qquad \phi : \mathbf{R}^d \to \mathbf{R}.$$

3. (Variance.) We have

$$E_f \max_{\phi\in\mathcal{D}} |\nu_n(\phi)| \le C\left(\frac{2\log_e(\#\mathcal{D})}{n}\right)^{1/2} + n^{-1/2}C'.$$

4. (Bias.) Since $\mathcal{F} = \bar{\text{co}}(\mathbb{G}) \cap \mathcal{F}_{den}$, and $\mathcal{D}$ is a finite δ-cover of $\mathbb{G}$, then

$$\sup_{f\in\mathcal{F}} \inf_{\Lambda\in\mathcal{W}} \|f - f(\cdot,\Lambda)\|_2 \le \delta.$$

Slide 15

An implementation

1. Choose $\tilde{f}_0$ as the greedy histogram constructed minimizing the L_2 empirical risk γ_n.
2. For $k = 1,\ldots,M-1$, let

$$\tilde{f}_k = (1-\pi_k)\tilde{f}_{k-1} + \pi_k\tilde{\phi}$$

where $\tilde{\phi}$ is the greedy histogram constructed minimizing the risk functional

$$\tilde{\gamma}_n(\phi) = \gamma_n\left(\pi_k\phi\right) + 2(1-\pi_k)\pi_k\frac{1}{B}\sum_{j=1}^B \phi(Z_j)$$

where $Z_1,\ldots,Z_B$ is a bootstrap sample from the distribution of $\tilde{f}_{k-1}$.

3. Let $\hat{f}_n = \tilde{f}_{M-1}$.

Slide 16

Explanation

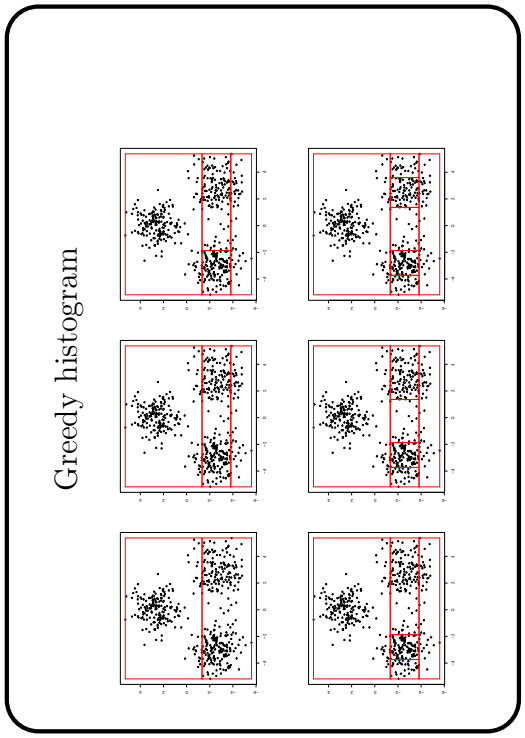
In stagewise minimization we minimize with respect to ϕ at each step

$$\begin{aligned}&\gamma_n\left((1-\pi_k)\hat{f}_{k-1} + \pi_k\phi\right)\\ &\quad= \gamma_n\left((1-\pi_k)\hat{f}_{k-1}\right) + \gamma_n\left(\pi_k\phi\right) + 2(1-\pi_k)\pi_k\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi.\end{aligned}$$

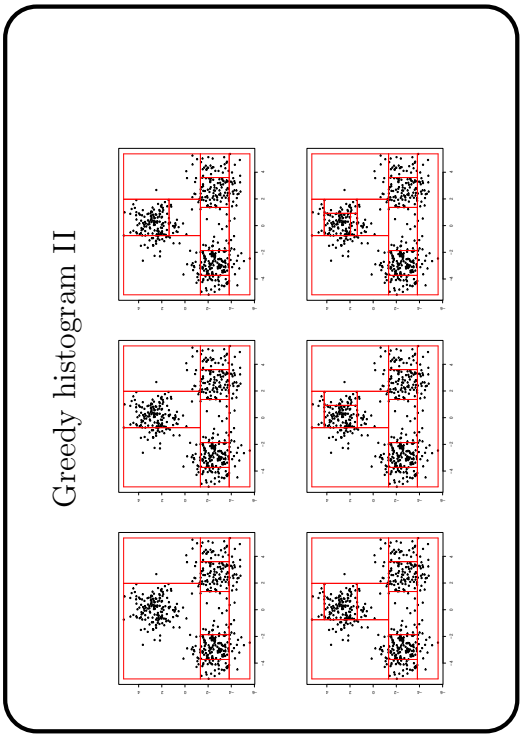
We may approximate the inner product with Monte Carlo integration:

$$\int_{\mathbf{R}^d}\hat{f}_{k-1}\phi \approx \frac{1}{B}\sum_{j=1}^B \phi(Z_j)$$

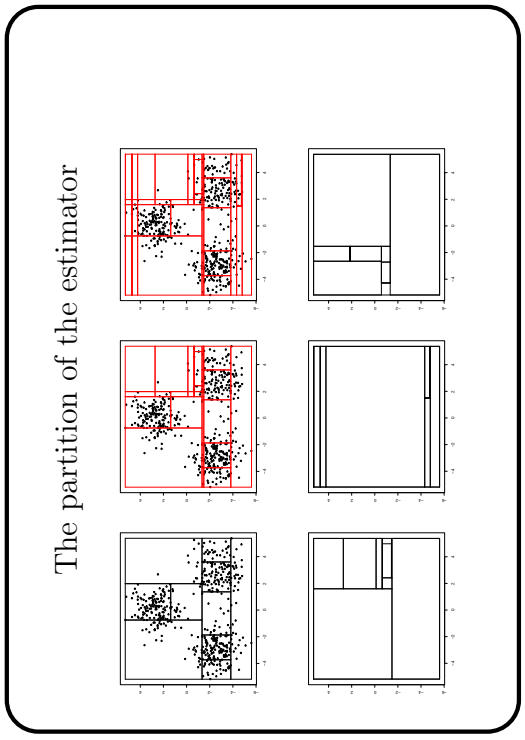
where $Z_1,\ldots,Z_B$ is a bootstrap sample from the distribution of $\hat{f}_{k-1}$.



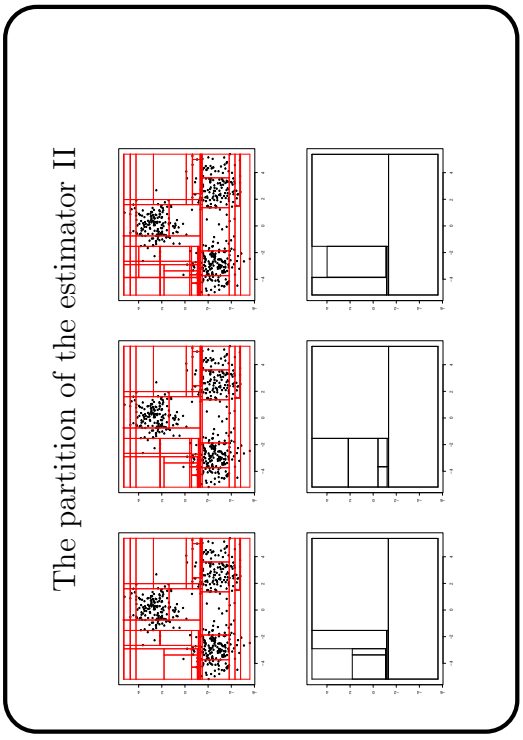
Slide 17



Slide 18



Slide 19



Slide 20