Model selection for fast density estimation

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

July 17, 2008

e-mail: gyorfi@szit.bme.hu
www.szit.bme.hu/˜gyorfi

Density estimation

$\mathbb{R}^d$-valued i.i.d. random vectors $X_1, \dots, X_n$

Density estimation

$\mathbb{R}^d$-valued i.i.d. random vectors $X_1, \ldots, X_n$
distributed according to unknown probability measure μ
with density f

Density estimation

$\mathbb{R}^d$-valued i.i.d. random vectors $X_1, \ldots, X_n$
distributed according to unknown probability measure μ
with density f
The L_1 norm

$$\|f - g\| := \int_{\mathbb{R}^d} |f(x) - g(x)| dx$$

Density estimation

$\mathbb{R}^d$-valued i.i.d. random vectors $X_1, \ldots, X_n$
distributed according to unknown probability measure μ
with density f
The L_1 norm

$$\|f - g\| := \int_{\mathbb{R}^d} |f(x) - g(x)| dx = 2 \sup_A \left| \int_A f(x) dx - \int_A g(x) dx \right|$$

Kernel density estimate

For a kernel function K and bandwidth $h > 0$,
let f_n be the kernel density estimate with sample size n:

$$f_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right).$$

Density-free consistency

If

$$\lim_{n\to\infty} h_n = 0$$

and

$$\lim_{n\to\infty} nh_n^d = \infty$$

Density-free consistency

If

$$\lim_{n\to\infty} h_n = 0$$

and

$$\lim_{n\to\infty} nh_n^d = \infty$$

then, for any density f,

$$\lim_{n\to\infty} \mathbf{E}\|f - f_n\| = 0$$

and

$$\lim_{n\to\infty} \|f - f_n\| = 0 \text{ a.s.}$$

Rate of convergence

If the density f has a compact support and is twice differentiable, then

$$\mathbf{E}(\|f_n - f\|) \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2.$$

Rate of convergence

If the density f has a compact support and is twice differentiable, then

$$\mathbf{E}(\|f_n - f\|) \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2.$$

If $h_n = cn^{-1/(d+4)}$ then

$$\mathbf{E}(\|f_n - f\|) \leq Cn^{-2/(d+4)}.$$

Rate of convergence

If the density f has a compact support and is twice differentiable, then

$$\mathbf{E}(\|f_n - f\|) \leq \frac{c_1}{\sqrt{nh_n^d}} + c_2 h_n^2.$$

If $h_n = cn^{-1/(d+4)}$ then

$$\mathbf{E}(\|f_n - f\|) \leq Cn^{-2/(d+4)}.$$

TOO SLOW.

Model selection for density estimation

We wish to estimate a density f on $\mathbb{R}^d$

Model selection for density estimation

We wish to estimate a density f on $\mathbb{R}^d$
that belongs to a parametric family, $\mathcal{F}_k$, where k is unknown,

Model selection for density estimation

We wish to estimate a density f on $\mathbb{R}^d$
that belongs to a parametric family, $\mathcal{F}_k$, where k is unknown,
but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k.

We wish to estimate a density f on $\mathbb{R}^d$
that belongs to a parametric family, $\mathcal{F}_k$, where k is unknown,
but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k.

$$\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k .$$

Model selection for density estimation

We wish to estimate a density f on $\mathbb{R}^d$
that belongs to a parametric family, $\mathcal{F}_k$, where k is unknown,
but $\mathcal{F}_k \subset \mathcal{F}_{k+1}$ for all k.

$$\mathcal{F} = \bigcup_{k \geq 1} \mathcal{F}_k.$$

the complexity associated with f is defined as

$$k^* = \min\{k \geq 1 : f \in \mathcal{F}_k\}.$$

Example

$$\mathcal{F}_k$$

is the set of mixtures of d dimensional normal densities, where the number of components is at most k

Objective

We wish to introduce an estimate k_n of the complexity k^* and

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \longrightarrow k^*$ almost surely

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \to k^*$ almost surely
   (i.e., $k_n = k^*$ almost surely, for all n large enough)

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \to k^*$ almost surely
   (i.e., $k_n = k^*$ almost surely, for all n large enough)
2. and
$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \to k^*$ almost surely
   (i.e., $k_n = k^*$ almost surely, for all n large enough)
2. and
$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Biau, Devroye (2004)

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \to k^*$ almost surely
   (i.e., $k_n = k^*$ almost surely, for all n large enough)
2. and
$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Biau, Devroye (2004)
k_n and $\hat{f}_{k_n}$ via projection of the empirical measure with respect to the Yatracos class

Objective

We wish to introduce an estimate k_n of the complexity k^* and to pick a density estimate $\hat{f}_{k_n}$ in $\mathcal{F}$ with

1. $k_n \to k^*$ almost surely
   (i.e., $k_n = k^*$ almost surely, for all n large enough)
2. and
$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Biau, Devroye (2004)
k_n and $\hat{f}_{k_n}$ via projection of the empirical measure with respect to the Yatracos class
too complex

Testing homogeneity

Testing homogeneity

Two mutually independent samples

$$X_1, \dots, X_n \quad \text{and} \quad X'_1, \dots, X'_n$$

distributed according to unknown probability distributions μ and μ' on $\mathbb{R}^d$.

Testing homogeneity

Two mutually independent samples

$$X_1, \ldots, X_n \quad \text{and} \quad X'_1, \ldots, X'_n$$

distributed according to unknown probability distributions μ and μ' on $\mathbb{R}^d$.
We are interested in testing the null hypothesis that the two samples are homogeneous, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

Testing homogeneity

Two mutually independent samples

$$X_1, \dots, X_n \quad \text{and} \quad X'_1, \dots, X'_n$$

distributed according to unknown probability distributions μ and μ' on $\mathbb{R}^d$.
We are interested in testing the null hypothesis that the two samples are homogeneous, that is

$$\mathcal{H}_0 : \mu = \mu'.$$

empirical probability distributions μ_n and μ'_n

The test statistic

Based on a partition $\mathcal{P}_n = \{A_{n1}, \ldots, A_{nm_n}\}$ of $\mathbb{R}^d$, we let the test statistic be defined as

$$T_n = \sum_{j=1}^{m_n} |\mu_n(A_{nj}) - \mu_n'(A_{nj})|.$$

Asymptotic behavior of T_n

Theorem. Under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,

$$\mathbb{P}\{T_n > \varepsilon\} = e^{-n(g_T(\varepsilon)+o(1))},$$

as $n \to \infty$,

Asymptotic behavior of T_n

Theorem. Under $\mathcal{H}_0$, for all $0 < \varepsilon < 2$,

$$\mathbb{P}\{T_n > \varepsilon\} = e^{-n(g_T(\varepsilon)+o(1))},$$

as $n \to \infty$, where

$$g_T(\varepsilon) = (1+\varepsilon/2)\ln(1+\varepsilon/2) + (1-\varepsilon/2)\ln(1-\varepsilon/2) \approx \varepsilon^2/4.$$

(Biau, Györfi (2005))

A strong consistent test

Corollary. Consider the test which rejects $\mathcal{H}_0$ when

$$T_n > 2\sqrt{\ln 2}\sqrt{\frac{m_n}{n}}.$$

A strong consistent test

Corollary. Consider the test which rejects $\mathcal{H}_0$ when

$$T_n > 2\sqrt{\ln 2}\sqrt{\frac{m_n}{n}}.$$

Assume that

$$\lim_{n\to\infty} \frac{m_n}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

A strong consistent test

Corollary. Consider the test which rejects $\mathcal{H}_0$ when

$$T_n > 2\sqrt{\ln 2}\sqrt{\frac{m_n}{n}}.$$

Assume that

$$\lim_{n\to\infty}\frac{m_n}{n}=0 \quad \text{and} \quad \lim_{n\to\infty}\frac{m_n}{\ln n}=\infty.$$

Then, under $\mathcal{H}_0$, after a random sample size the test makes a.s. no error.

A strong consistent test

Corollary. Consider the test which rejects $\mathcal{H}_0$ when

$$T_n > 2\sqrt{\ln 2}\sqrt{\frac{m_n}{n}}.$$

Assume that

$$\lim_{n\to\infty} \frac{m_n}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

Then, under $\mathcal{H}_0$, after a random sample size the test makes a.s. no error.

Moreover, if $\mu \neq \mu'$, and for each sphere S centered at the origin

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0$$

A strong consistent test

Corollary. Consider the test which rejects $\mathcal{H}_0$ when

$$T_n > 2\sqrt{\ln 2}\sqrt{\frac{m_n}{n}}.$$

Assume that

$$\lim_{n\to\infty} \frac{m_n}{n} = 0 \quad \text{and} \quad \lim_{n\to\infty} \frac{m_n}{\ln n} = \infty.$$

Then, under $\mathcal{H}_0$, after a random sample size the test makes a.s. no error.
Moreover, if $\mu \neq \mu'$, and for each sphere S centered at the origin

$$\lim_{n\to\infty} \max_{j:A_{n,j}\cap S\neq\emptyset} diam(A_{n,j}) = 0$$

then after a random sample size the test makes a.s. no error.
(Biau, Györfi (2005))

Complexity estimation

Split the sample into two subsamples:

$$\{X_1, \ldots, X_n\} \quad \text{and} \quad \{X'_1, \ldots, X'_n\} = \{X_{n+1}, \ldots, X_{2n}\}.$$

Complexity estimation

Split the sample into two subsamples:

$$\{X_1, \ldots, X_n\} \quad \text{and} \quad \{X'_1, \ldots, X'_n\} = \{X_{n+1}, \ldots, X_{2n}\}.$$

Let $\mathcal{P}_n = \{A_{nj} : j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d.

Complexity estimation

Split the sample into two subsamples:

$$\{X_1, \ldots, X_n\} \quad \text{and} \quad \{X'_1, \ldots, X'_n\} = \{X_{n+1}, \ldots, X_{2n}\}.$$

Let $\mathcal{P}_n = \{A_{nj} : j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d. Introduce the statistic

$$d_{n,k} = \inf_{g \in \mathcal{F}_k} \sum_{A \in \mathcal{P}_n} \left| \int_A g - \mu_{2n}(A) \right|.$$

Complexity estimation

Split the sample into two subsamples:

$$\{X_1, \dots, X_n\} \quad \text{and} \quad \{X'_1, \dots, X'_n\} = \{X_{n+1}, \dots, X_{2n}\}.$$

Let $\mathcal{P}_n = \{A_{nj} : j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d. Introduce the statistic

$$d_{n,k} = \inf_{g \in \mathcal{F}_k} \sum_{A \in \mathcal{P}_n} \left| \int_A g - \mu_{2n}(A) \right|.$$

Let the threshold be

$$T_n = \sum_{A \in \mathcal{P}_n} |\mu_n(A) - \mu'_n(A)|.$$

Split the sample into two subsamples:

$$\{X_1, \dots, X_n\} \quad \text{and} \quad \{X'_1, \dots, X'_n\} = \{X_{n+1}, \dots, X_{2n}\}.$$

Let $\mathcal{P}_n = \{A_{nj} : j \geq 1\}$ be a cubic partition of $\mathbb{R}^d$ with volume h_n^d. Introduce the statistic

$$d_{n,k} = \inf_{g \in \mathcal{F}_k} \sum_{A \in \mathcal{P}_n} \left| \int_A g - \mu_{2n}(A) \right|.$$

Let the threshold be

$$T_n = \sum_{A \in \mathcal{P}_n} |\mu_n(A) - \mu'_n(A)|.$$

Estimate of k^*:

$$k_n = \min\{k \geq 1 : d_{n,k} \leq T_n\}.$$

Theorem 1

Assume that, for each $k \geq 1$, $\mathcal{F}_k$ is closed with respect to the weak convergence topology.

Assume that, for each $k \geq 1$, $\mathcal{F}_k$ is closed with respect to the weak convergence topology.
Then there exists a positive constant κ, depending on f, such that

$$\mathbb{P}\left\{k_n \neq k^*\right\} \leq \exp\left(-\kappa\, h_n^{-d}\right),$$

Theorem 1

Assume that, for each $k \geq 1$, $\mathcal{F}_k$ is closed with respect to the weak convergence topology.
Then there exists a positive constant κ, depending on f, such that

$$\mathbb{P}\left\{k_n \neq k^*\right\} \leq \exp\left(-\kappa\, h_n^{-d}\right),$$

and consequently, for the choice $h_n = n^{-\delta}$ with $0 < \delta < 1/d$,

$$k_n = k^*$$

almost surely, for all n large enough.
(Biau, Cadre, Devroye, Györfi (2008))

Fix $k \geq 1$ and introduce the (Yatracos) class of sets

$$\mathcal{A}_k = \left\{ \{x : g_1(x) > g_2(x)\} : g_1, g_2 \in \mathcal{F}_k \right\}$$

Fast density estimate

Fix $k \geq 1$ and introduce the (Yatracos) class of sets

$$\mathcal{A}_k = \left\{ \{x : g_1(x) > g_2(x)\} : g_1, g_2 \in \mathcal{F}_k \right\}$$

and the goodness criterion for a density $g \in \mathcal{F}_k$:

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_{2n}(A) \right|.$$

Fast density estimate

Fix $k \geq 1$ and introduce the (Yatracos) class of sets

$$\mathcal{A}_k = \left\{ \{x : g_1(x) > g_2(x)\} : g_1, g_2 \in \mathcal{F}_k \right\}$$

and the goodness criterion for a density $g \in \mathcal{F}_k$:

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_{2n}(A) \right|.$$

The minimum distance estimate $\hat{f}_k$ minimizes the criterion $\Delta_k(g)$ over all g in $\mathcal{F}_k$.

Fast density estimate

Fix $k \geq 1$ and introduce the (Yatracos) class of sets

$$\mathcal{A}_k = \left\{ \{x : g_1(x) > g_2(x)\} : g_1, g_2 \in \mathcal{F}_k \right\}$$

and the goodness criterion for a density $g \in \mathcal{F}_k$:

$$\Delta_k(g) = \sup_{A \in \mathcal{A}_k} \left| \int_A g - \mu_{2n}(A) \right|.$$

The minimum distance estimate $\hat{f}_k$ minimizes the criterion $\Delta_k(g)$ over all g in $\mathcal{F}_k$.
The density estimate is

$$\hat{f}_{k_n}.$$

Theorem 2

If $\mathcal{A}_{k^*}$ has finite Vapnik-Chervonenkis dimension

Theorem 2

If $\mathcal{A}_{k^*}$ has finite Vapnik-Chervonenkis dimension then

$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

Theorem 2

If $\mathcal{A}_{k^*}$ has finite Vapnik-Chervonenkis dimension then

$$\mathbb{E}\left\{\|\hat{f}_{k_n} - f\|\right\} = O\left(\frac{1}{\sqrt{n}}\right).$$

(Biau, Devroye (2004))

Problem

The projection with respect to the Yatracos class is too complex.

Problem

The projection with respect to the Yatracos class is too complex.
For a kernel function K and bandwidth $r > 0$,
let f_{2n} be the kernel density estimate with sample size $2n$:

$$f_{2n}(x) = \frac{1}{2nr^d} \sum_{i=1}^{2n} K\left(\frac{x - X_i}{r}\right).$$

Problem

The projection with respect to the Yatracos class is too complex.
For a kernel function K and bandwidth $r > 0$,
let f_{2n} be the kernel density estimate with sample size $2n$:

$$f_{2n}(x) = \frac{1}{2nr^d} \sum_{i=1}^{2n} K\left(\frac{x - X_i}{r}\right).$$

let $K_r * g$ be the expectation of the kernel estimate with density g:

$$K_r * g(x) = \frac{1}{r^d} \int K\left(\frac{x - z}{r}\right) g(z) dz.$$

Problem

The projection with respect to the Yatracos class is too complex.
For a kernel function K and bandwidth $r > 0$,
let f_{2n} be the kernel density estimate with sample size $2n$:

$$f_{2n}(x) = \frac{1}{2nr^d} \sum_{i=1}^{2n} K\left(\frac{x - X_i}{r}\right).$$

let $K_r * g$ be the expectation of the kernel estimate with density g:

$$K_r * g(x) = \frac{1}{r^d} \int K\left(\frac{x - z}{r}\right) g(z)dz.$$

the estimate $\bar{f}_n$ is defined as

$$\bar{f}_n = \arg\min_{g \in \mathcal{F}_{k_n}} \|K_r * g - f_{2n}\|,$$

Problem

The projection with respect to the Yatracos class is too complex.
For a kernel function K and bandwidth $r > 0$,
let f_{2n} be the kernel density estimate with sample size $2n$:

$$f_{2n}(x) = \frac{1}{2nr^d} \sum_{i=1}^{2n} K\left(\frac{x - X_i}{r}\right).$$

let $K_r * g$ be the expectation of the kernel estimate with density g:

$$K_r * g(x) = \frac{1}{r^d} \int K\left(\frac{x - z}{r}\right) g(z)dz.$$

the estimate $\bar{f}_n$ is defined as

$$\bar{f}_n = \arg\min_{g \in \mathcal{F}_{k_n}} \|K_r * g - f_{2n}\|,$$

$\bar{f}_n$ is an L_1-projection of the kernel density estimate f_{2n}
with fixed bandwidth r.

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before
such that the bandwidth is $h = h_n = (\ln n)^{-(1+\delta)/d}$ with $\delta > 0$

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before
such that the bandwidth is $h = h_n = (\ln n)^{-(1+\delta)/d}$ with $\delta > 0$
Choose the kernel function K such that it is a density function and its characteristic function is everywhere non-zero.

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before
such that the bandwidth is $h = h_n = (\ln n)^{-(1+\delta)/d}$ with $\delta > 0$
Choose the kernel function K such that it is a density function and its characteristic function is everywhere non-zero.
Suppose that

$$\sup_{g \in \mathcal{F}_{k^*}} \frac{\|g - f\|}{\|K_r * g - K_r * f\|} < \infty,$$

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before
such that the bandwidth is $h = h_n = (\ln n)^{-(1+\delta)/d}$ with $\delta > 0$
Choose the kernel function K such that it is a density function and its characteristic function is everywhere non-zero.
Suppose that

$$\sup_{g \in \mathcal{F}_{k^*}} \frac{\|g - f\|}{\|K_r * g - K_r * f\|} < \infty,$$

and

$$\int \sqrt{f} < \infty.$$

Theorem 3

Assume that $\mathcal{F}_k$ is closed in the weak convergence topology for every $k \geq 1$.
Choose k_n as before
such that the bandwidth is $h = h_n = (\ln n)^{-(1+\delta)/d}$ with $\delta > 0$
Choose the kernel function K such that it is a density function and its characteristic function is everywhere non-zero.
Suppose that

$$\sup_{g \in \mathcal{F}_{k^*}} \frac{\|g - f\|}{\|K_r * g - K_r * f\|} < \infty,$$

and

$$\int \sqrt{f} < \infty.$$

Then

$$\mathbf{E}\left\{\|\bar{f}_n - f\|\right\} \leq O\left(\frac{1}{\sqrt{n}}\right).$$

Luc's problem

Let f be the density of a multidimensional normal distribution.

Luc's problem

Let f be the density of a multidimensional normal distribution. Find the optimal density estimate in L_1.

Luc's problem

Let f be the density of a multidimensional normal distribution. Find the optimal density estimate in L_1.

$$\min_{f_n} \mathbf{E}\|f_n - f\|$$

Luc's problem

Let f be the density of a multidimensional normal distribution. Find the optimal density estimate in L_1.

$$\min_{f_n} \mathbf{E}\|f_n - f\|$$

The plug-in estimate is not optimal.