A combined strategy for multivariate density estimation.

Alejandro Cholaquidis

CMAT-Facultad de Ciencias, UdelaR
Montevideo Uruguay

Joint work with: R. Fraiman, B. Ghattas and J. Kalemkerian

Seminario de Probabilidad y Estadística

1 The general setup
- An alternative approach
- Optimality

2 A Central limit theorem

3 Simulations

1 The general setup
- An alternative approach
- Optimality

2 A Central limit theorem

3 Simulations

The idea

Our approach is based on two main ideas:

1) Compute the estimator of f at x using

$$\{y : |f(y) - f(x)| \leq \epsilon\} \equiv B^*(\epsilon, x),$$

instead of a neighborhood of x,

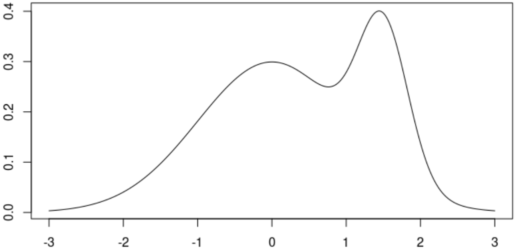

Figure: Left: a density whose concentration mass varies significantly in its support. Right: the 0.2-neighborhood for the level $f(x) = 0.2$ is given by the union of the intervals $I_1 = [-1.01, -0.78]$ and $I_2 = [1.86, 1.92]$

The idea

Our approach is based on two main ideas:

1) Compute the estimator of f at x using

$$\{y : |f(y) - f(x)| \leq \epsilon\} \equiv B^*(\epsilon, x),$$

instead of a neighborhood of x,

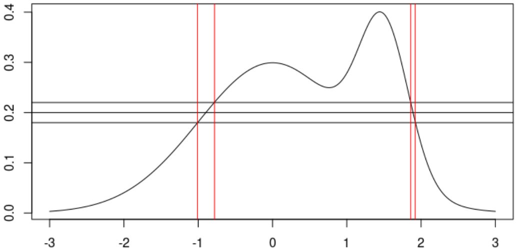

Figure: Left: a density whose concentration mass varies significantly in its support. Right: the 0.2-neighborhood for the level $f(x) = 0.2$ is given by the union of the intervals $I_1 = [-1.01, -0.78]$ and $I_2 = [1.86, 1.92]$

The idea

Our approach is based on two main ideas:

1) Compute the estimator of f at x using

$$\{y : |f(y) - f(x)| \leq \epsilon\} \equiv B^*(\epsilon, x),$$

instead of a neighborhood of x,

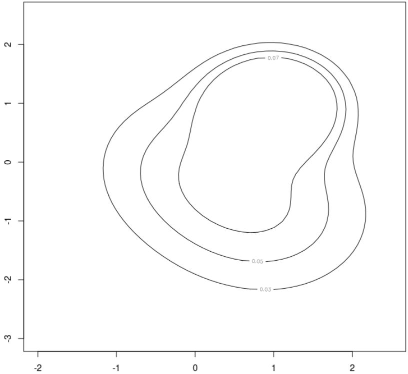

Figure: $B(0.02$,-neighbourhood for a mixture of three bi-variate gaussian distributions: one with mean $(0, 0)$ and covariance matrix $\Sigma = diag(1, 1)$, the second one with mean $(1, 1)$ and covariance matrix $\Sigma = diag(1/2, 1/2)$, the last one with mean $(1, -1)$ and covariance matrix $\Sigma = diag(1, 1)$, $f(x) = 0.05$

The idea

2) Perform a nonlinear aggregation method to combine several estimators. This will improve the behavior when for instance the underling true density f is not unimodal, and the concentration of mass varies significantly within its support.

The idea

2) Perform a nonlinear aggregation method to combine several estimators. This will improve the behavior when for instance the underling true density f is not unimodal, and the concentration of mass varies significantly within its support.

Notation

1) $f : S \subset \mathbb{R}^d \to \mathbb{R}$ is a density bounded from above and $\int f^2(x)dx < \infty$

The idea

2) Perform a nonlinear aggregation method to combine several estimators. This will improve the behavior when for instance the underling true density f is not unimodal, and the concentration of mass varies significantly within its support.

Notation

1) $f : S \subset \mathbb{R}^d \to \mathbb{R}$ is a density bounded from above and $\int f^2(x)dx < \infty$

2) $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ iid from X with density f. We split $\mathcal{D}_n$ into two disjoint subsets, namely $\mathcal{D}_k = \{X_1, \ldots, X_k\}$ and $\mathcal{E}_l = \{X_{k+1}, \ldots, X_n\}$ with $l = n - k$.

The idea

2) Perform a nonlinear aggregation method to combine several estimators. This will improve the behavior when for instance the underling true density f is not unimodal, and the concentration of mass varies significantly within its support.

Notation

1) $f : S \subset \mathbb{R}^d \to \mathbb{R}$ is a density bounded from above and $\int f^2(x)dx < \infty$

2) $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ iid from X with density f. We split $\mathcal{D}_n$ into two disjoint subsets, namely $\mathcal{D}_k = \{X_1, \ldots, X_k\}$ and $\mathcal{E}_l = \{X_{k+1}, \ldots, X_n\}$ with $l = n - k$.

3) $\mathbf{f_k}(x) = (f_1(x), \ldots, f_M(x))$ density estimators computed with $\mathcal{D}_k$.

The idea

2) Perform a nonlinear aggregation method to combine several estimators. This will improve the behavior when for instance the underling true density f is not unimodal, and the concentration of mass varies significantly within its support.

Notation

1) $f : S \subset \mathbb{R}^d \to \mathbb{R}$ is a density bounded from above and $\int f^2(x)dx < \infty$

2) $\mathcal{D}_n = \{X_1, \ldots, X_n\}$ iid from X with density f. We split $\mathcal{D}_n$ into two disjoint subsets, namely $\mathcal{D}_k = \{X_1, \ldots, X_k\}$ and $\mathcal{E}_l = \{X_{k+1}, \ldots, X_n\}$ with $l = n - k$.

3) $\mathbf{f_k}(x) = (f_1(x), \ldots, f_M(x))$ density estimators computed with $\mathcal{D}_k$.

The estimator

Let $\epsilon > 0$ and $x \in \mathbb{R}^d$, define

$$B(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \bigcap_{m=1}^{M} |f_m(y) - f_m(x)| < \epsilon \right\}.$$

$$N(\epsilon, x) = \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x))$$

The estimator

Let $\epsilon > 0$ and $x \in \mathbb{R}^d$, define

$$B(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \bigcap_{m=1}^{M} |f_m(y) - f_m(x)| < \epsilon \right\}.$$

$$N(\epsilon, x) = \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x))$$

Intuitively, observe that

$$\frac{P_X(B(\epsilon, x))}{\mu(B(\epsilon, x))} = \frac{1}{\mu(B(\epsilon, x))} \int_{B(\epsilon, x)} f(x)dx \sim f(x),$$

The estimator

Let $\epsilon > 0$ and $x \in \mathbb{R}^d$, define

$$B(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \bigcap_{m=1}^{M} |f_m(y) - f_m(x)| < \epsilon \right\}.$$

$$N(\epsilon, x) = \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x))$$

Intuitively, observe that

$$\frac{P_X(B(\epsilon, x))}{\mu(B(\epsilon, x))} = \frac{1}{\mu(B(\epsilon, x))} \int_{B(\epsilon, x)} f(x) dx \sim f(x),$$

and for l large enough

$$P_X(B(\epsilon, x)) \sim \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x)).$$

The estimator

Let $\epsilon > 0$ and $x \in \mathbb{R}^d$, define

$$B(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \bigcap_{m=1}^{M} |f_m(y) - f_m(x)| < \epsilon \right\}.$$

$$N(\epsilon, x) = \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x))$$

Intuitively, observe that

$$\frac{P_X(B(\epsilon, x))}{\mu(B(\epsilon, x))} = \frac{1}{\mu(B(\epsilon, x))} \int_{B(\epsilon, x)} f(x)dx \sim f(x),$$

and for l large enough

$$P_X(B(\epsilon, x)) \sim \frac{1}{l} \#(\mathcal{E}_l \cap B(\epsilon, x)).$$

The aggregated density estimator is defined as

$$\hat{f}_{\text{agg}}(x) = \frac{N(\epsilon, x)}{\mu(B(\epsilon, x))}$$

Let $\epsilon > 0$ and $0 \leq \eta < 1$,

$$B^{\eta}(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}_{\{|f_m(y) - f_m(x)| < \epsilon\}} \geq 1 - \eta \right\}.$$

Let $\epsilon > 0$ and $0 \leq \eta < 1$,

$$B^{\eta}(\epsilon, x) = \left\{ y \in \mathbb{R}^d : \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}_{\{|f_m(y) - f_m(x)| < \epsilon\}} \geq 1 - \eta \right\}.$$

For $\eta = 0$ we get $B^{\eta}(\epsilon, x) = B(\epsilon, x)$.

Define the η-density estimator, $\hat{f}_{\text{agg},\eta}(x)$ as

$$\tilde{f}_{\text{agg},\eta}(x) = \frac{1}{\mu(B^{\eta}(\epsilon, x)) l} \#(\mathcal{E}_l \cap B^{\eta}(\epsilon, x)).$$

Given X independent of $\mathcal{D}_n$, let us define,

$$T(\mathbf{f_k}(X)) = \mathbb{E}(f(X)|\mathbf{f_k}(X)).$$

Proposition

$$\mathbb{E}|\hat{f}_{\text{agg}}(X) - f(X)|^2 \leq \min_{m=1,\dots,M} \mathbb{E}|f_m(X) - f(X)|^2 + \mathbb{E}|\hat{f}_{\text{agg}}(X) - T(\mathbf{f_k}(X))|^2$$

Given X independent of $\mathcal{D}_n$, let us define,

$$T(\mathbf{f_k}(X)) = \mathbb{E}(f(X)|\mathbf{f_k}(X)).$$

Proposition

$$\mathbb{E}|\hat{f}_{\text{agg}}(X) - f(X)|^2 \leq \min_{m=1,\ldots,M} \mathbb{E}|f_m(X) - f(X)|^2 + \mathbb{E}|\hat{f}_{\text{agg}}(X) - T(\mathbf{f_k}(X))|^2$$

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i \to \infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i \to \infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i \to \infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

- $f_m(x) \to f(x)$ for $m = 1, \ldots, M$, a.s., as $k \to \infty$.

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i\to\infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

- $f_m(x) \to f(x)$ for $m = 1, \ldots, M$, a.s., as $k \to \infty$.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$.

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i\to\infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

- $f_m(x) \to f(x)$ for $m = 1, \ldots, M$, a.s., as $k \to \infty$.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$.
- $\overline{B^*(\epsilon, x)}$ is compact, and $\overline{B(\epsilon, x)}$ is compact a.s.

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i\to\infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

- $f_m(x) \to f(x)$ for $m = 1, \ldots, M$, a.s., as $k \to \infty$.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$.
- $\overline{B^*(\epsilon, x)}$ is compact, and $\overline{B(\epsilon, x)}$ is compact a.s.

we have

$$\mu(B(\epsilon, x)) \to \mu(B^*(\epsilon, x)) \quad a.s., \text{ as } k \to \infty, \tag{1}$$

and

K1 A random variable X with distribution P_X and density f fulfils $K1$, if $\mathbb{P}(f(X) = a) = 0$ for all $a \in \mathbb{R}$.

Lemma

Let us assume that $K1$ holds. Let f_i be a density estimator of f such $f_i(X) \to f(X)$ a.s, as $i \to \infty$. Then

$$\lim_{i \to \infty} \mathbb{E}\left|\mathbb{E}[f(X)|f_i(X)] - f(X)\right|^2 = 0.$$

Lemma

Let X be random variable with distribution P_X whose density f is continuous. Let be $f_1, \ldots, f_M$ continuous, such that for all $m = 1, \ldots, M$, $|f_m(x) - f(x)| \to 0$ a.s., as $k \to \infty$ for almost all x w.r.t to μ. Let $\epsilon > 0$, then for all x such that

- $f_m(x) \to f(x)$ for $m = 1, \ldots, M$, a.s., as $k \to \infty$.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$.
- $\overline{B^*(\epsilon, x)}$ is compact, and $\overline{B(\epsilon, x)}$ is compact a.s.

we have

$$\mu(B(\epsilon, x)) \to \mu(B^*(\epsilon, x)) \quad a.s., \text{ as } k \to \infty, \tag{1}$$

and

$$P_X(B(\epsilon, x)) \to P_X(B^*(\epsilon, x)) \quad a.s., \text{ as } k \to \infty. \tag{2}$$

We will consider the following set of assumptions

We will consider the following set of assumptions

H1 The density estimators $f_1, \ldots, f_M$ based on a sample $\mathcal{D}_k$ fulfils H1 if with probability one, the sequences $\{f_1\}_k, \ldots, \{f_M\}_k$ are uniformly equicontinuous and the $\delta = \delta(\epsilon)$ of the uniform equicontinuity is bounded from below by $\delta_0(\epsilon) > 0$.

We will consider the following set of assumptions

H1 The density estimators $f_1, \ldots, f_M$ based on a sample $\mathcal{D}_k$ fulfils H1 if with probability one, the sequences $\{f_1\}_k, \ldots, \{f_M\}_k$ are uniformly equicontinuous and the $\delta = \delta(\epsilon)$ of the uniform equicontinuity is bounded from below by $\delta_0(\epsilon) > 0$.

H2 The density estimators $f_1, \ldots, f_M$ based on a sample $\mathcal{D}_k$ fulfils H2 if for almost all x w.r.t. μ, $f_j(x) \rightarrow f(x)$, a.s., for all $j = 1, \ldots, M$ as $k \rightarrow \infty$.

We will consider the following set of assumptions

- H1 The density estimators $f_1, \ldots, f_M$ based on a sample $\mathcal{D}_k$ fulfils H1 if with probability one, the sequences $\{f_1\}_k, \ldots, \{f_M\}_k$ are uniformly equicontinuous and the $\delta = \delta(\epsilon)$ of the uniform equicontinuity is bounded from below by $\delta_0(\epsilon) > 0$.
- H2 The density estimators $f_1, \ldots, f_M$ based on a sample $\mathcal{D}_k$ fulfils H2 if for almost all x w.r.t. μ, $f_j(x) \to f(x)$, a.s., for all $j = 1, \ldots, M$ as $k \to \infty$.

Theorem

Let us assume K1, H1 and H2. Assume also that,

- for all x such that $f_m(x) \to f(x)$ for all $m = 1, \ldots, M$ there exists $\epsilon_0(x)$ such that for all $0 < \epsilon < \epsilon_0(x)$, the set $\overline{B^*(\epsilon, x)}$ is compact and the set $\overline{B(\epsilon, x)}$ is compact a.s.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$.

Let $k = k(l) \to \infty$ as $l \to \infty$, then

$$\lim_{\epsilon \to 0} \lim_{l \to \infty} \mathbb{E}|\hat{f}_{agg}(X) - T(\mathbf{f_k}(X))|^2 = 0. \quad (3)$$

1 The general setup
- An alternative approach
- Optimality

2 A Central limit theorem

3 Simulations

Let us denote $B^*(\epsilon, x) = \{y : |f(x) - f(y)| < \epsilon\}$.

Theorem

Let $\epsilon = \epsilon_l \to 0$ such that $l\epsilon_l^2 \to 0$. Then, for all x such that $f(x) > 0$ and

- $\mu(\{y : f(x) = f(y)\}) = 0$
- $\overline{\mu(B^*(\epsilon, x))}$ is compact, and $\overline{B(\epsilon, x)}$ is compact a.s.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$
- $\mu(B^*(\epsilon, x))l \to \infty$
- $f_m(x) \to f(x)$ for all $m = 1, \dots, M$.

We have,

$$\lim_{l\to\infty} \lim_{k\to\infty} \sqrt{\mu(B^*(\epsilon, x))l}\left[\hat{f}_{agg}(x) - f(x)\right] \stackrel{d}{=} N(0, f(x)). \quad (4)$$

Let us denote $B^*(\epsilon, x) = \{y : |f(x) - f(y)| < \epsilon\}$.

Theorem

Let $\epsilon = \epsilon_l \to 0$ such that $l\epsilon_l^2 \to 0$. Then, for all x such that $f(x) > 0$ and

- $\mu(\{y : f(x) = f(y)\}) = 0$
- $\overline{\mu(B^*(\epsilon, x))}$ is compact, and $\overline{B(\epsilon, x)}$ is compact a.s.
- $\mu[B^*(\epsilon + \gamma, x) \setminus B^*(\epsilon - \gamma, x)] \to 0$ as $\gamma \to 0$
- $\mu(B^*(\epsilon, x))l \to \infty$
- $f_m(x) \to f(x)$ for all $m = 1, \ldots, M$.

We have,

$$\lim_{l\to\infty} \lim_{k\to\infty} \sqrt{\mu(B^*(\epsilon, x))l}\left[\hat{f}_{agg}(x) - f(x)\right] \stackrel{d}{=} N(0, f(x)). \qquad (4)$$

Proposition

Let f be a spherical density (i.e., $f(x) = h(\|x\|^2)$ for some $h : \mathbb{R} \to \mathbb{R}$) such that h is strictly decreasing and h' is continuous on a neighbourhood containing $\|x\|^2$, then, for all x such that $f(x) > 0$, and $\|\nabla f(x)\| > 0$,

$$\lim_{l\to\infty} \frac{\mu(B^*(\epsilon, x))}{2\epsilon} = \frac{2\pi^{d/2}\|x\|^{d-1}}{\Gamma(\frac{d}{2})\|\nabla f(x)\|},$$

where Γ is the Euler's gamma function.

1 The general setup
- An alternative approach
- Optimality

2 A Central limit theorem

3 Simulations

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix.

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix.
- Weibull, whose density is given by $\left(\frac{k}{\lambda^k}\right)^d (x_1 \cdots x_d)^{d(k-1)} \exp\left(-\sum_{i=1}^d (x_i/\lambda)^k\right)$,

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ is a diagonal matrix.
- Weibull, whose density is given by $\left(\frac{k}{\lambda^k}\right)^d (x_1 \cdots x_d)^{d(k-1)} \exp\left(-\sum_{i=1}^d (x_i/\lambda)^k\right)$,

To build $\hat{f}_{\text{agg}}$ considered 5 kernels $f_{k,\gamma_1}, \dots, f_{k,\gamma_5}$ computed with $\gamma_1, \dots, \gamma_5$ where:

- First we compute LOO *hcv* based on a sample of size k. This value is kept fixed along the replicates.

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1} \cdots (1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ is a diagonal matrix.
- Weibull, whose density is given by $\left(\frac{k}{\lambda^k}\right)^d (x_1 \cdots x_d)^{d(k-1)} \exp\left(-\sum_{i=1}^d (x_i/\lambda)^k\right)$,

To build $\hat{f}_{\text{agg}}$ considered 5 kernels $f_{k,\gamma_1}, \dots, f_{k,\gamma_5}$ computed with $\gamma_1, \dots, \gamma_5$ where:

- First we compute LOO *hcv* based on a sample of size k. This value is kept fixed along the replicates.
- Fix $\gamma_1 = 0.9hcv$, $\gamma_2 = 0.95hcv$, $\gamma_3 = hcv$, $\gamma_4 = 1.05hcv$ and $\gamma_5 = 1.1hcv$.

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1 \cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ is a diagonal matrix.
- Weibull, whose density is given by $\left(\frac{k}{\lambda^k}\right)^d (x_1 \cdots x_d)^{d(k-1)} \exp\left(-\sum_{i=1}^d (x_i/\lambda)^k\right)$,

To build $\hat{f}_{\text{agg}}$ considered 5 kernels $f_{k,\gamma_1}, \ldots, f_{k,\gamma_5}$ computed with $\gamma_1, \ldots, \gamma_5$ where:

- First we compute LOO *hcv* based on a sample of size k. This value is kept fixed along the replicates.
- Fix $\gamma_1 = 0.9hcv$, $\gamma_2 = 0.95hcv$, $\gamma_3 = hcv$, $\gamma_4 = 1.05hcv$ and $\gamma_5 = 1.1hcv$.
- In general we took $k = l = 2000$ for $d = 2$ and $k = l = 4000$ for $d = 4$.

Three different distributions were considered:

- Beta, given by $\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^d (x_1\cdots x_d)^{\alpha-1}(1-x_1)^{\beta-1}\cdots(1-x_d)^{\beta-1}$.
- Normal, with mean 0 and variance $\Sigma = \mathrm{diag}(\sigma_1^2,\dots,\sigma_d^2)$ is a diagonal matrix.
- Weibull, whose density is given by $\left(\frac{k}{\lambda^k}\right)^d (x_1\cdots x_d)^{d(k-1)} \exp\left(-\sum_{i=1}^d (x_i/\lambda)^k\right)$,

To build $\hat{f}_{\mathrm{agg}}$ considered 5 kernels $f_{k,\gamma_1},\dots,f_{k,\gamma_5}$ computed with $\gamma_1,\dots,\gamma_5$ where:

- First we compute LOO *hcv* based on a sample of size k. This value is kept fixed along the replicates.
- Fix $\gamma_1 = 0.9hcv$, $\gamma_2 = 0.95hcv$, $\gamma_3 = hcv$, $\gamma_4 = 1.05hcv$ and $\gamma_5 = 1.1hcv$.
- In general we took $k = l = 2000$ for $d = 2$ and $k = l = 4000$ for $d = 4$.
- Denote *hcvu* the LOO based on $\mathcal{D}_k \cup \mathcal{E}_l$.
  ϵ_l was selected as follows: compute $f_{k+l,\tilde{h}_1},\dots,f_{k+l,\tilde{h}_5}$ based on $\mathcal{D}_k \cup \mathcal{E}_l$, where $\tilde{h}_1 = 0.9hcvu$, $\tilde{h}_2 = 0.95hcvu$, $\tilde{h}_3 = hcvu$, $\tilde{h}_4 = 1.05hcvu$ and $\tilde{h}_5 = 1.1hcvu$, define

$$\bar{f}(x) = \frac{f_{k+l,\tilde{h}_1} + \dots + f_{k+l,\tilde{h}_5}}{5}.$$

  Finally $\epsilon_l = \mathrm{argmin}\|\hat{f}_{\mathrm{agg}} - \bar{f}\|_2$.

	$\alpha = 1.5, \beta = 1.5$			$\alpha = 2.5, \beta = 2.5$	
	$d = 2$,	$d = 4$,		$d = 2$	$d = 4$
n, k	2000	4000		2000	4000
Kernel	G	G		G	G
$\hat{f}_{agg}$	**0.080**	**0.179**		0.122	0.325
$f_{k,0.9*hcv}$	0.113	0.260		0.126	0.385
$f_{k,0.95*hcv}$	0.113	0.262		0.125	0.364
$f_{k,hcv}$	0.114	0.267		0.125	0.350
$f_{k,1.05*hcv}$	0.116	0.274		0.126	0.343
$f_{k,1.1*hcv}$	0.118	0.282		0.128	0.343
$f_{k+l,0.9*hcv}$	0.094	0.236		0.101	**0.249**
$f_{k+l,0.95*hcv}$	0.096	0.244		0.103	0.252
$f_{k+l,hcvu}$	0.099	0.236		**0.101**	0.279
$f_{k+l,1.05*hcv}$	0.103	0.263		0.109	0.271
$f_{k+l,1.1*hcv}$	0.118	0.283		0.129	0.343

	$\alpha = 1.5, \beta = 1.5$			$\alpha = 2.5, \beta = 2.5$	
	$d = 2$,	$d = 4$,		$d = 2$	$d = 4$
k, l	2000	4000		2000	4000
Kernel	E	E		E	E
$\hat{f}_{agg}$	**0.065**	**0.131**		**0.147**	0.240
$f_{k,0.9*hcv}$	0.089	0.174		0.267	0.319
$f_{k,0.95*hcv}$	0.089	0.166		0.258	0.293
$f_{k,hcv}$	0.092	0.163		0.250	0.272
$f_{k,1.05*hcv}$	0.099	0.162		0.242	0.256
$f_{k,1.1*hcv}$	0.108	0.165		0.235	0.244
$f_{k+l,0.9*hcv}$	0.085	0.146		0.244	0.265
$f_{k+l,0.95*hcv}$	0.085	0.144		0.238	0.246
$f_{k+l,hcvu}$	0.089	0.146		0.222	**0.220**
$f_{k+l,1.05*hcv}$	0.096	0.147		0.225	0.271
$f_{k+l,1.1*hcv}$	0.108	0.165		0.235	0.244

Table: L_2 error over 100 repetitions with beta distributions.

	$\lambda = 1, k = 1$			$\lambda = 1, k = 0.5$	
	$d = 2,$	$d = 4,$		$d = 2$	$d = 4$
n, k Kernel	2000 E	4000 G		2000 E	4000 G
$\hat{f}_{\text{agg}}$	0.035	0.065		**0.139**	**0.054**
$f_{k,0.9*hcv}$	0.041	**0.064**		0.184	0.083
$f_{k,0.95*hcv}$	0.039	0.066		0.183	0.083
$f_{k,hcv}$	0.038	0.067		0.182	0.083
$f_{k,1.05*hcv}$	0.036	0.068		0.182	0.084
$f_{k,1.1*hcv}$	0.035	0.069		0.181	0.086
$f_{k+l,0.9*hcv}$	0.035	0.064		0.179	0.083
$f_{k+l,0.95*hcv}$	0.034	0.065		0.179	0.082
$f_{k+l,hcvu}$	0.034	0.065		0.180	0.082
$f_{k+l,1.05*hcv}$	**0.032**	0.068		0.178	0.084
$f_{k+l,1.1*hcv}$	0.035	0.069		0.181	0.086

Table: L_2 error over 100 repetitions with Weibull distributions.

	d=2	d=2	$d = 4$
	$\sigma_1 = 1, \sigma_2 = 0.4$	$\sigma_1 = 1, \sigma_2 = 0.1$	$\sigma_1 = .1 = \sigma_2$, $\sigma_3 = 1 = \sigma_4$
$n = k$ Kernel	2000 E	2000 E	4000 E
$\hat{f}_{agg}$	**0.0130**	**0.0328**	**0.065**
$f_{k,0.9*hcv}$	0.0162	0.0425	0.083
$f_{k,0.95*hcv}$	0.0163	0.0418	0.086
$f_{k,hcv}$	0.0164	0.0415	0.087
$f_{k,1.05*hcv}$	0.0169	0.0416	0.089
$f_{k,1.1*hcv}$	0.0174	0.0420	0.091
$f_{k+l,0.9*hcv}$	0.0154	0.0373	0.083
$f_{k+l,0.95*hcv}$	0.0156	0.0372	0.085
$f_{k+l,hcvu}$	0.0154	0.0374	0.087
$f_{k+l,1.05*hcv}$	0.0164	0.0379	0.089
$f_{k+l,1.1*hcv}$	0.0175	0.0420	0.091

Table: L_2 error for the Normal distribution over 100 repetitions using Epanechnikov's kernel. In $\mathbb{R}^2$ $\Sigma = diag(\sigma_1^2, \sigma_2^2)$, and in $\mathbb{R}^4$ $\Sigma = diag(\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)$.

- DEVROYE, L., GYÖRFI, L. AND LUGOSI, G. (1996).
  A Probabilistic Theory of Pattern Recognition. Springer-Verlag, New York.
- Agrawala, A.K. (1970). Learning with a probabilistic teacher.
  IEEE Transactions on Automatic Control, (19), 716–723
- Belkin, M. and Niyogi, P. (2004). Semi-supervised learning on Riemannian manifolds.
  Machine Learning **56** pp. 209–239.
- Ben-David, S., Lu, T. and Pal, D.(2008). Does unlabelled data provably help?. Worst-case analysis of the sample complexity of semi-supervised learning.
  In 21*st Annual Conference on Learning Theory (COLT)*. Available at `http://www.informatik.uni-trier.de/~ley/db/conf/colt/colt2008.html`.
- Chapelle, O., Schölkopf, B. and Zien, A., eds. (2006) Semi-supervised learning.
  Adaptative computation and machine learning series. MIT
- Fralick, S.C. (1967) Learning recognize patterns without teacher.
  IEEE Transactions on Information Theory (13) 57–64.
- Haffari, G. and Sarkar, A. (2007) Analysis of Semi-Supervised Learning with the Yarkowsky algoritm.
  In Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence, UAI 2007. Vancouver, BC. July 19-22, 2007.
- Zhu, X. (2008) Semi-supervised learning literature survey.
  `http://pages.cs.wisc.edu/~jerryzhu/research/ssl/semireview.html`