A tailor made nonparametric density estimate

Daniel Carando [1], Ricardo Fraiman [2] and Pablo Groisman [1]

[1]Universidad de Buenos Aires

[2]Universidad de San Andrés

School and Workshop on Probability Theory and Applications ICTP/CNPq
January 2007, Campos do Jordão, Brasil

The density estimation problem

- X a random variable on $\mathbb{R}^d$ with density f.
- The density f is unknown.
- We have an i.i.d. sample $X_1, \ldots, X_n$ drawn from f.

The density estimation problem

- X a random variable on $\mathbb{R}^d$ with density f.
- The density f is unknown.
- We have an i.i.d. sample $X_1, \ldots, X_n$ drawn from f.

We look for an estimate of f based in the sample. That is, a mapping $f_n \colon \mathbb{R}^d \times (\mathbb{R}^d)^n \to \mathbb{R}$

The density estimation problem

- X a random variable on $\mathbb{R}^d$ with density f.
- The density f is unknown.
- We have an i.i.d. sample $X_1, \dots, X_n$ drawn from f.

We look for an estimate of f based in the sample. That is, a mapping $f_n \colon \mathbb{R}^d \times (\mathbb{R}^d)^n \to \mathbb{R}$

The density f is assumed to belong to a certain class $\mathcal{F}$.

Maximum likelihood estimates

For every density function g

$$\mathbb{E}(\log g(X)) \leq \mathbb{E}(\log f(X))$$

Maximum likelihood estimates

For every density function g

$$\mathbb{E}(\log g(X)) \leq \mathbb{E}(\log f(X))$$

The Maximum Likelihood Estimate (MLE) is defined as the maximizer of the empirical mean (log-likelihood function)

$$\mathcal{L}_n(g) := \frac{1}{n}\sum_{i=1}^{n} \log g(X_i) = \log\left(\prod_{i=1}^{n} g(X_i)\right)^{1/n}$$

over the class $\mathcal{F}$

Maximum likelihood estimates

For every density function g

$$\mathbb{E}(\log g(X)) \leq \mathbb{E}(\log f(X))$$

The Maximum Likelihood Estimate (MLE) is defined as the maximizer of the empirical mean (log-likelihood function)

$$\mathcal{L}_n(g) := \frac{1}{n} \sum_{i=1}^{n} \log g(X_i) = \log \left(\prod_{i=1}^{n} g(X_i) \right)^{1/n}$$

over the class $\mathcal{F}$

This problem is not always well posed (depending on the class $\mathcal{F}$)

The parametric case

The class $\mathcal{F}$ can be parameterized (by a finite number of parameters)

Example: $X \sim N(\mu, \sigma^2)$

$$\mathcal{F} = \{f_{(\mu,\sigma^2)}, \mu \in \mathbb{R}, \sigma^2 > 0\}$$

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The maximizer f_n is the density of a gaussian random variable with parameters

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_n)^2.$$

The parametric case

Under some (fairly weak) conditions, in the parametric case, the MLE are known to be

1. Strongly consistent, i.e. $f_n \longrightarrow f$ a.s.
2. Asymptotically minimum variance unbiased estimators
3. Asymptotically gaussian.

The MLE make use of the knowledge we have on f since depends strongly on the class $\mathcal{F}$ where we look for the maximizer

The nonparametric case

The class $\mathcal{F}$ has infinite dimension

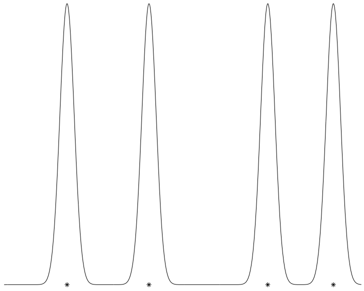
Example: $\mathcal{F} = \{g \in L^1(\mathbb{R}^d), g \geq 0, \|g\|_{L^1} = 1\}$, the class of all densities.

The nonparametric case

The class $\mathcal{F}$ has infinite dimension

Example: $\mathcal{F} = \{g \in L^1(\mathbb{R}^d), g \geq 0, \|g\|_{L^1} = 1\}$, the class of all densities.

The MLE method fails since $\mathcal{L}_n(g)$ is unbounded



Approximations of the identity belong to $\mathcal{F}$

Alternatives

Kernel density estimates (Parzen and Rosenblatt, 1956)

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

$K \geq 0,\ \int K = 1,\ h > 0$

Alternatives

Kernel density estimates (Parzen and Rosenblatt, 1956)

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

$K \geq 0$, $\int K = 1$, $h > 0$

Very popular. Very flexible.

Alternatives

Kernel density estimates (Parzen and Rosenblatt, 1956)

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

$K \geq 0$, $\int K = 1$, $h > 0$

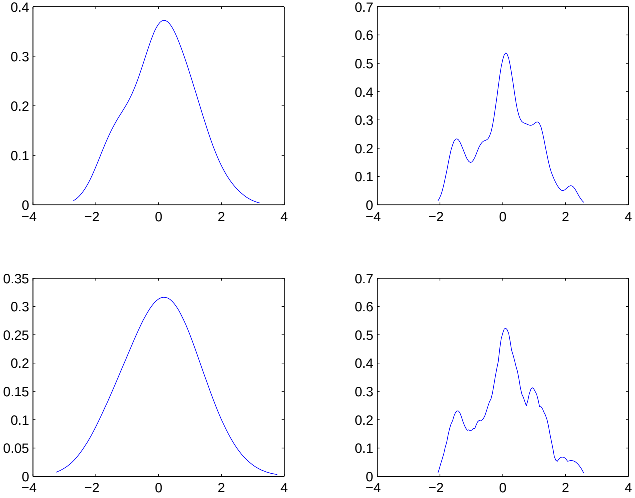
Very popular. Very flexible.

The kernel K and the bandwidth h must be chosen

Alternatives

The kernel density estimate

The kernel density estimate for different choices of K and h



The kernel density estimate is a “Universal” estimate. It works for all densities f. Does not make use of further knowledge on f.

Alternatives

Maximum penalized likelihood estimates, Good and Gaskins (1971)

Idea: Penalize the lack of smoothness

Instead of looking for a maximizer of $\mathcal{L}_n(g)$, we look for a maximizer of

$$\frac{1}{n}\left(\sum_{i=1}^{n}\log g(X_i) - h\int g''^2\right).$$

Alternatives

Tailor-designed Maximum Likelihood Estimates

If we have some knowledge on f then $\mathcal{F}$ is not the class of all densities and, may be, we can apply MLE techniques

- Grenander (1956) considered $\mathcal{F}$ to be the class of decreasing densities in $\mathbb{R}_+$

- Grenander (1956) considered $\mathcal{F}$ to be the class of decreasing densities in $\mathbb{R}_+$
- In this case it turns out that the MLE is well defined and is the derivative of the least concave majorant of the empirical distribution function

Tailor-designed Maximum Likelihood Estimates

Grenander's estimate

- Grenander (1956) considered $\mathcal{F}$ to be the class of decreasing densities in $\mathbb{R}_+$
- In this case it turns out that the MLE is well defined and is the derivative of the least concave majorant of the empirical distribution function
- It is consistent and minimax optimal in this class.

Tailor-designed Maximum Likelihood Estimates

Grenander's estimate

- Grenander (1956) considered $\mathcal{F}$ to be the class of decreasing densities in $\mathbb{R}_+$
- In this case it turns out that the MLE is well defined and is the derivative of the least concave majorant of the empirical distribution function
- It is consistent and minimax optimal in this class.
- Robertson (1967), Wegman (1969, 1970), Sager (1982) and Polonik (1998) generalized Grenander's estimate to other kinds of "shape restrictions"

Tailor-designed Maximum Likelihood Estimates

Grenander's estimate

- Grenander (1956) considered $\mathcal{F}$ to be the class of decreasing densities in $\mathbb{R}_+$
- In this case it turns out that the MLE is well defined and is the derivative of the least concave majorant of the empirical distribution function
- It is consistent and minimax optimal in this class.
- Robertson (1967), Wegman (1969, 1970), Sager (1982) and Polonik (1998) generalized Grenander's estimate to other kinds of "shape restrictions"

MLE for Lipschitz densities

We consider $\mathcal{F}$ to be the class of densities g with compact support $S(g)$ that verify

$$|g(x) - g(y)| \leq \kappa \|x - y\|, \qquad x, y \in S(g).$$

That is, $\mathcal{F}$ is the class of Lipschitz densities with prescribed Lipschitz constant κ. We allow g to be discontinuous at the boundary of its support.

MLE for Lipschitz densities

We consider $\mathcal{F}$ to be the class of densities g with compact support $S(g)$ that verify

$$|g(x) - g(y)| \leq \kappa \|x - y\|, \qquad x, y \in S(g).$$

That is, $\mathcal{F}$ is the class of Lipschitz densities with prescribed Lipschitz constant κ. We allow g to be discontinuous at the boundary of its support.

The support of the density f can be unknown
(In this case we ask $S(f)$ to be convex)

Theorem

(i) *There exists a unique maximizer* f_n *of* $\mathcal{L}_n(g)$ *in* $\mathcal{F}$*. Moreover,* f_n *is supported in* $\mathcal{C}_n$*, the convex hull of* $\{X_1, \ldots, X_n\}$*, and its value there is given by the maximum of* n *"cone functions", i.e.*

$$f_n(x) = \max_{1 \le i \le n} \left(f_n(X_i) - \kappa \|x - X_i\|\right)^+ . \qquad (1)$$

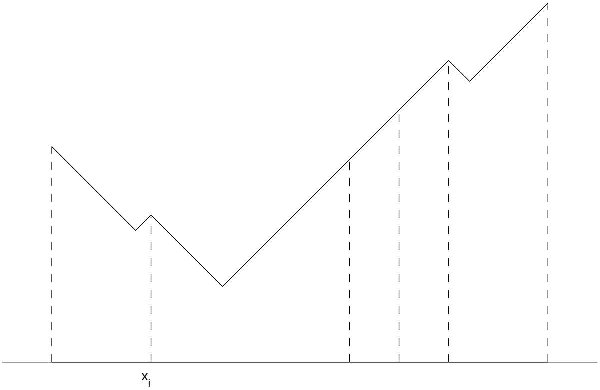
(ii) f_n *is consistent in the following sense: for every compact set* $K \subset S(f)^\circ$,

$$\lim_{n\to\infty} \|f_n - f\|_{L^\infty(K)} \to 0 \qquad a.s.$$

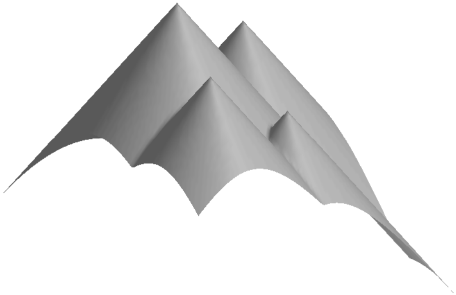
(iii) *Hence*

$$\lim_{n\to\infty} \|f_n - f\|_{L^1(\mathbb{R}^d)} \to 0 \qquad a.s.$$

The MLE in dimension $d = 1$



The MLE in dimension $d = 2$



Proof.

(i) Existence $\rightarrow$ Picture. Uniqueness $\rightarrow$ We are looking for a maximum of a concave function in a convex set.

Proof.

(i) Existence $\to$ Picture. Uniqueness $\to$ We are looking for a maximum of a concave function in a convex set.

(ii) Is a consequence of Huber's Theorem (1967).

Idea:
Use Huber's Theorem we need a sequence $\hat{f}_n$ of (almost) maximizers of $\mathcal{L}_n$ belonging to a (fixed) compact class.
We construct them us follows

$$\hat{f}_n := A_n \max_{1 \leq i \leq n} \left(f_n(X_i) - \kappa \|x - X_i\| \right)^+ , \qquad \text{for all } x \in S(f).$$

The constant A_n is chosen to guarantee $\int f_n = 1$.

Proof.

(i) Existence $\rightarrow$ Picture. Uniqueness $\rightarrow$ We are looking for a maximum of a concave function in a convex set.

(ii) Is a consequence of Huber's Theorem (1967).

Idea:
Use Huber's Theorem we need a sequence $\hat{f}_n$ of (almost) maximizers of $\mathcal{L}_n$ belonging to a (fixed) compact class.
We construct them us follows

$$\hat{f}_n := A_n \max_{1\leq i\leq n} \left(f_n(X_i) - \kappa\|x - X_i\|\right)^+ , \qquad \text{for all } x \in S(f).$$

The constant A_n is chosen to guarantee $\int f_n = 1$.

And $\hat{f}_n \in \mathrm{Lip}(\kappa, S(f))$, which is compact

$$\|f_n - \hat{f}_n\|_{L^\infty(K)} \leq |A_n - 1| \; \|f_n\|_{L^\infty(K)} \rightarrow 0, \tag{2}$$

since $A_n \rightarrow 1$ and $(\|f_n\|_{L^\infty(K)})_n$ is bounded a.s.

(iii) Since

- $\mathcal{C}_n \subset S(f)$
- $|S(f)| < \infty$
- $|f_n(x)| \leq \kappa \mathrm{diam}(S(f)) + \frac{1}{|\mathcal{C}_n|}$

we can find $K \subset S(f)$ such that

$$\int_{\mathbb{R}^d} |f_n(x) - f(x)|\, dx \leq$$

$$\int_K |f_n(x) - f(x)|\, dx + \int_{S(f)\setminus K} |f_n(x) - f(x)|\, dx \to \varepsilon$$

Computing the estimator

We have proved that the estimator lives in a certain finite-dimensional space and that is determined by its value at the sample points.

For $y \in \mathbb{R}^n$ we define

$$g_y(x) = \max_{1 \le i \le n} \Big(y_i - |x - X_i| \Big)^+, \qquad x \in \mathcal{C}_n.$$

Our problem read us

Find

$$\mathrm{argmax}_{y \in \mathcal{P}} \prod_{i=1}^{n} y_i.$$

$$\mathcal{P} = \{ y \in \mathbb{R}^n, y_i > 0, |y_i - y_j| \le \kappa |X_i - X_j| (i \ne j), \int g_y = 1 \}.$$

$\mathcal{P}$ is convex and $\prod y_i$ is concave

To have an efficient method to solve this problem we need to decide (efficiently) if a point $y \in \mathcal{P}$

Easy in $d = 1$. Not so easy if $d > 1$

Computing the estimator

Dimension $d = 1$

Let $(X^{(1)}, \ldots, X^{(n)})$ the order statistics. The Lipschitz conditions reads us

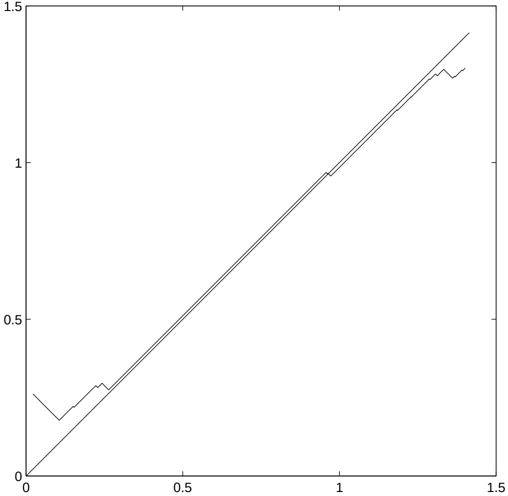
$$-\kappa(X^{(i+1)} - X^{(i)}) \leq y_{i+1} - y_i \leq \kappa(X^{(i+1)} - X^{(i)}), \qquad i = 1, \ldots, n-1.$$

And $\displaystyle\int g_y(x)\, dx =$

$$= \frac{1}{4} \sum_{i=1}^{n-1} (y_{i+1} - y_i)^2 + 2(y_{i+1} + y_i)(X^{(i+1)} - X^{(i)}) - (X^{(i+1)} - X^{(i)})^2.$$

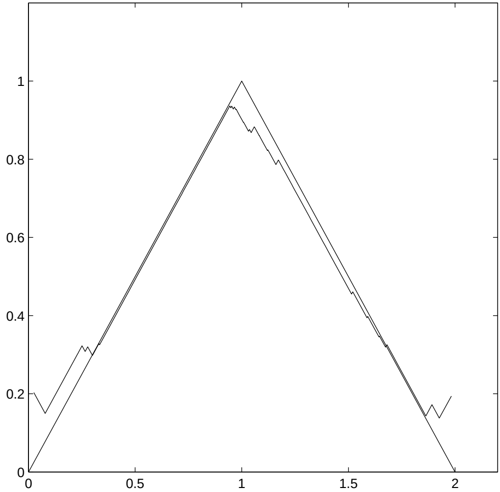
Computing the estimator

Dimension $d=1$ - Sample size: $n=100$.



Computing the estimator

Dimension $d = 1$ - Sample size: $n = 100$.



Computing the estimator

Dimension $d > 1$

- We can not order the sample points
- We have not an explicit formula for the integral $\int g_y(x)\, dx$

Some problems...

- Too many peaks

Some problems...

- Too many peaks
- An optimization nonlinear problem has to be solved.

An alternative ML type estimator

Dimension one - PLMLE

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) =$$

$$\left\{ g \in \mathrm{Lip}(\kappa, [X^{(1)}, X^{(n)}]) : g|_{[X^{(i)}, X^{(i+1)}]} \text{ is linear } \int g = 1 \right\},$$

An alternative ML type estimator

Dimension one - PLMLE

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) =$$

$$\left\{ g \in \mathrm{Lip}(\kappa, [X^{(1)}, X^{(n)}]) : g|_{[X^{(i)}, X^{(i+1)}]} \text{ is linear } \int g = 1 \right\},$$

Definition

The PLMLE is the maximizer $\tilde{f}_n$ of $\mathcal{L}_n$ over $\mathcal{V}(X_1, \ldots, X_n)$.

An alternative ML type estimator

Dimension one - PLMLE

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) =$$

$$\left\{ g \in \mathrm{Lip}(\kappa, [X^{(1)}, X^{(n)}]) : g|_{[X^{(i)}, X^{(i+1)}]} \text{ is linear } \int g = 1 \right\},$$

Definition

The PLMLE is the maximizer $\tilde{f}_n$ of $\mathcal{L}_n$ over $\mathcal{V}(X_1, \ldots, X_n)$.

Existence and uniqueness of this estimator is guaranteed since $\mathcal{V}$ is a finite dimensional compact and convex subset of $\mathcal{F}$.

An alternative ML type estimator

Dimension one - PLMLE

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) =$$

$$\left\{ g \in \mathrm{Lip}(\kappa, [X^{(1)}, X^{(n)}]) : g|_{[X^{(i)}, X^{(i+1)}]} \text{ is linear } \int g = 1 \right\},$$

Definition

The PLMLE is the maximizer $\tilde{f}_n$ of $\mathcal{L}_n$ over $\mathcal{V}(X_1, \ldots, X_n)$.

Existence and uniqueness of this estimator is guaranteed since $\mathcal{V}$ is a finite dimensional compact and convex subset of $\mathcal{F}$.

It has lower likelihood than f_n but is asymptotically the same

Computation of PLMLE

maximize $\prod_{i=1}^{n} y_i$; subject to

$$-a \le Ay \le a, \qquad By = 1.$$

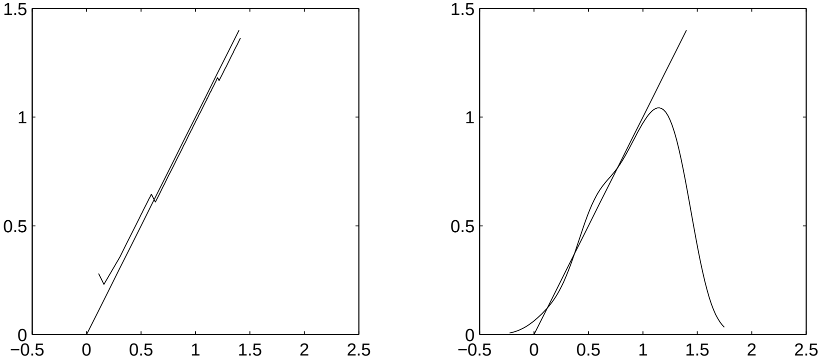
$$A = \begin{pmatrix} -1 & 1 & 0 & \cdots & & & 0 \\ 0 & -1 & 1 & & & & \\ \vdots & & \ddots & \ddots & & & \vdots \\ & & & -1 & & 1 & 0 \\ 0 & \cdots & & 0 & & -1 & 1 \end{pmatrix}, \qquad a = \kappa \begin{pmatrix} x_2 - x_1 \\ \vdots \\ x_{i+1} - x_i \\ \vdots \\ x_n - x_{n-1} \end{pmatrix},$$

$$B = \frac{1}{2}(x_2 - x_1, x_3 - x_1, \ldots, x_{i+1} - x_{i-1}, \ldots, x_n - x_{n-2}, x_n - x_{n-1})$$

The equation $-a \le Ay \le a$ guarantees the Lipschitz condition and $By = 1$ represents the restriction $\int \tilde{f}_n = 1$.

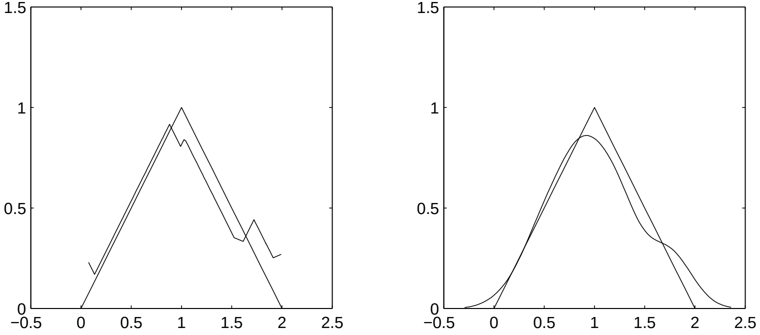
PLMLE demonstration

PLMLE vs. Kernels. Sample size: n=100

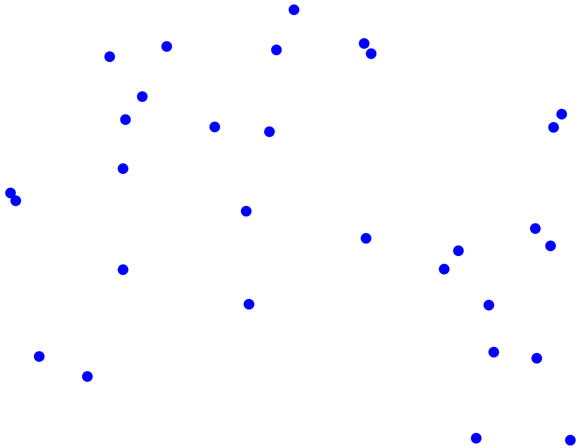


PLMLE demonstration

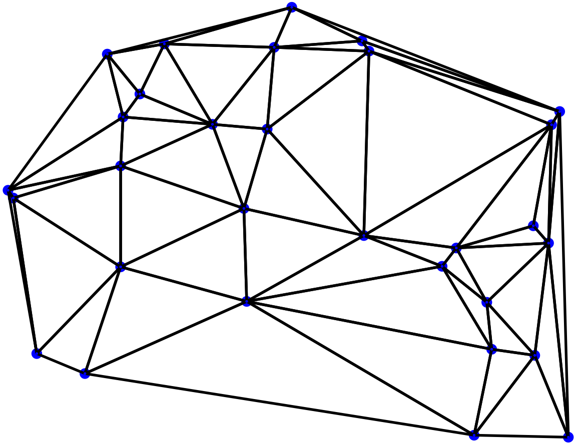
PLMLE vs. Kernels. Sample size: n=100



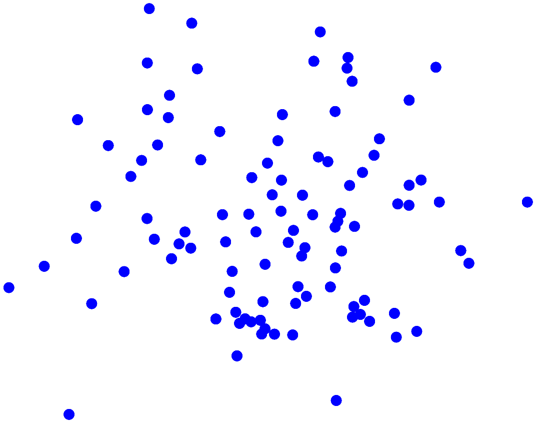
Delaunay triangulations



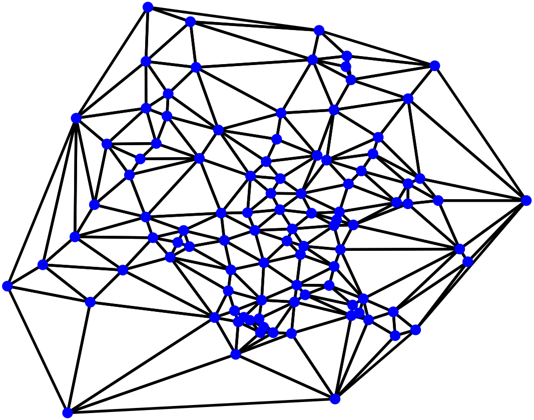
Delaunay triangulations



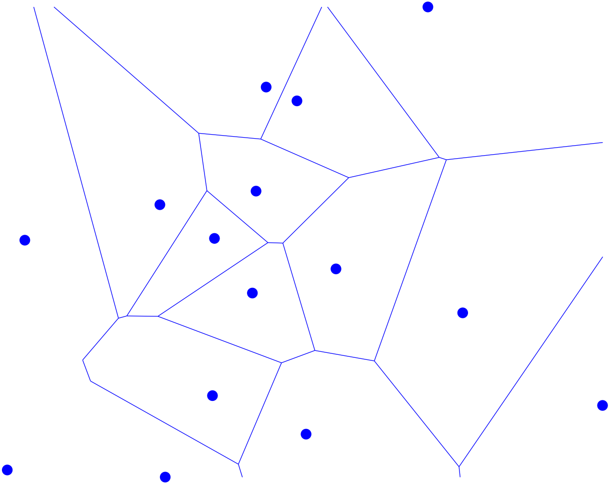
Delaunay triangulations



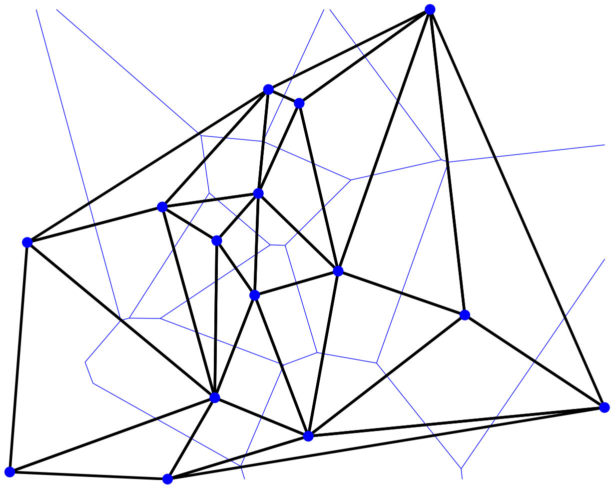
Delaunay triangulations



Voronoi tessellations and Delaunay triangulations



Voronoi tessellations and Delaunay triangulations



Some facts on Delaunay Triangulations

- For sample points of absolutely continuous probabilities is well defined with probability one.

Some facts on Delaunay Triangulations

- For sample points of absolutely continuous probabilities is well defined with probability one.
- Maximizes the mimimum angle of the triangles among all possible triangulations of the points.

Some facts on Delaunay Triangulations

- For sample points of absolutely continuous probabilities is well defined with probability one.
- Maximizes the mimimum angle of the triangles among all possible triangulations of the points.
- Useful for the numerical treatment of Partial Differential Equations by the Finite Element Method.

Some facts on Delaunay Triangulations

- For sample points of absolutely continuous probabilities is well defined with probability one.
- Maximizes the mimimum angle of the triangles among all possible triangulations of the points.
- Useful for the numerical treatment of Partial Differential Equations by the Finite Element Method.
- Useful to compute the Euclidean Minimum Spanning Tree of a set of points (Is a subgraph of the Delaunay triangulation).

Some facts on Delaunay Triangulations

- For sample points of absolutely continuous probabilities is well defined with probability one.
- Maximizes the mimimum angle of the triangles among all possible triangulations of the points.
- Useful for the numerical treatment of Partial Differential Equations by the Finite Element Method.
- Useful to compute the Euclidean Minimum Spanning Tree of a set of points (Is a subgraph of the Delaunay triangulation).
- Very used in Computational Geometry.

$T = \{\tau_1, \ldots, \tau_N\}$ The Delaunay Tesselation.

$T = \{\tau_1, \ldots, \tau_N\}$ The Delaunay Tesselation.

$$\mathcal{C}_n = \bigcup_{i=1}^{N} \tau_i$$

$T = \{\tau_1, \ldots, \tau_N\}$ The Delaunay Tesselation.

$$\mathcal{C}_n = \bigcup_{i=1}^{N} \tau_i$$

For any $i \neq j$, $\tau_i \cap \tau_j$ is either a point, a $(d-1)-$dimensional face, or the empty set.

We consider now the class of piecewise linear functions on T

$T = \{\tau_1, \ldots, \tau_N\}$ The Delaunay Tesselation.

$$\mathcal{C}_n = \bigcup_{i=1}^{N} \tau_i$$

For any $i \neq j$, $\tau_i \cap \tau_j$ is either a point, a $(d-1)-$dimensional face, or the empty set.

We consider now the class of piecewise linear functions on T

$$\mathcal{V} = \mathcal{V}(X_1, \ldots, X_n) =$$

$$\left\{ g \in \mathrm{Lip}(\kappa, \mathcal{C}_n) : g|_{\tau_i} \text{ is linear}, \int g = 1 \right\},$$

Definition

The PLMLE $\tilde{f}_n$ is the argument that maximizes $\mathcal{L}_n$ over $\mathcal{V}(X_1, \ldots, X_n)$.

Theorem

For every compact set $K \subset S(f)$ *we have*

$$\|\tilde{f}_n - f\|_{L^\infty(K)} \to 0 \qquad a.s.$$

Computing the estimator

$\mathcal{V}$ is a compact subset of the (finite dimensional) vector space

$$\tilde{\mathcal{V}} = \{g : \mathcal{C}_n \to \mathbb{R} : g|_{\tau_i} \text{ is linear}\}$$

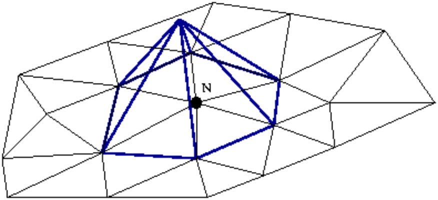
We need a (good) basis for $\tilde{\mathcal{V}}$.

Computing the estimator

$\mathcal{V}$ is a compact subset of the (finite dimensional) vector space

$$\tilde{\mathcal{V}} = \{g : \mathcal{C}_n \to \mathbb{R} : g|_{\tau_i} \text{ is linear}\}$$

We need a (good) basis for $\tilde{\mathcal{V}}$. We borrow from FEM.



$$g \in \tilde{\mathcal{V}} \Rightarrow$$

$$g(x) = \sum_i g(X_i)\varphi_i(x)$$

$$\varphi_i(X_j) = \delta_{ij}.$$

Computing the estimator

$$\int_{\mathbb{R}^d} g(x)\,dx = \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} g(X_i)\varphi_i(x) \right) dx = By,$$

Computing the estimator

$$\int_{\mathbb{R}^d} g(x)\,dx = \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} g(X_i)\varphi_i(x) \right) dx = By,$$

$$B = \left(\int \varphi_1, \ldots, \int \varphi_n \right) \qquad y = (g(X_1), \ldots, g(X_n))$$

Computing the estimator

$$\int_{\mathbb{R}^d} g(x)\,dx = \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} g(X_i)\varphi_i(x) \right) dx = By,$$

$$B = \left(\int \varphi_1, \ldots, \int \varphi_n \right) \qquad y = (g(X_1), \ldots, g(X_n))$$

We compute B just once!

Computing the estimator

$$\int_{\mathbb{R}^d} g(x)\,dx = \int_{\mathbb{R}^d} \left(\sum_{i=1}^{n} g(X_i)\varphi_i(x) \right) dx = By,$$

$$B = \left(\int \varphi_1, \ldots, \int \varphi_n \right) \qquad y = (g(X_1), \ldots, g(X_n))$$

We compute B just once!

We also have

$$\nabla g|_{\tau_k} = \sum_i y_i \nabla \varphi_i|_{\tau_k} = A_k y,$$

$$A_k = \left((\nabla \varphi_1|_{\tau_k})^t \middle| \cdots \middle| (\nabla \varphi_n|_{\tau_k})^t \right),$$

Computing the estimator

The optimization problem reads us

maximize $\prod_{i=1}^{n} y_i$; subject to

$$\|A_k y\| \leq \kappa, \quad 1 \leq k \leq N, \qquad By = 1.$$

Computing the estimator

The optimization problem reads us

maximize $\prod_{i=1}^{n} y_i$; subject to

$$\|A_k y\| \leq \kappa, \quad 1 \leq k \leq N, \qquad By = 1.$$

- If $\|\cdot\| = \|\cdot\|_\infty$, all the restrictions are linear.

Computing the estimator

The optimization problem reads us

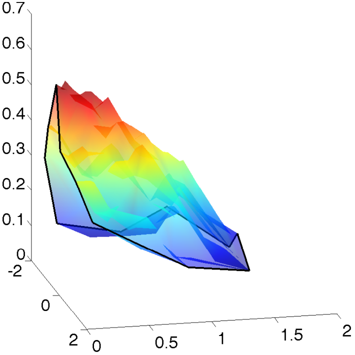
maximize $\prod_{i=1}^{n} y_i$; subject to

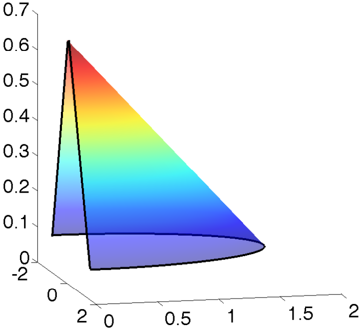
$$\|A_k y\| \leq \kappa, \quad 1 \leq k \leq N, \qquad By = 1.$$

- If $\| \cdot \| = \| \cdot \|_\infty$, all the restrictions are linear.
- The size of A_k grows linearly with d

The PLMLE for a "cone" density

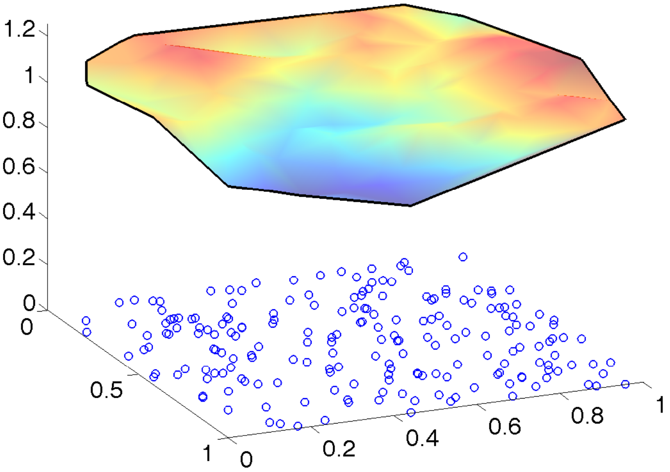
Sample size: n=250





The PLMLE for a Uniform density

Sample size: n=200



The PLMLE for a bivariate sum of uniform variables

Sample size: n=400