Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence

Colloque Jeunes Probabilistes et Statisticiens

Richard Fischer

EDF R&D MRI, CERMICS, LAMA
Supervisors: Cristina Butucea (LAMA), Jean-François Delmas (CERMICS),
Anne Dutfoy (EDF R&D MRI)

18/04/2016

Summary

Summary

Estimation problem

- Suppose that we have an i.i.d. sample $\mathcal{X}^n = (X^1, X^2, \ldots, X^n)$ of a d-dimensional distribution whose density has a product form on $\triangle = \{x = (x_1, \ldots, x_d) \in \mathbb{R}^d : 0 \leq x_1 \leq x_2 \leq \ldots \leq x_d \leq 1\}$:

$$f(x) = \prod_{i=1}^{d} p_i(x_i) \mathbf{1}_{\triangle}(x) = \mathrm{e}^{\left(\sum_{i=1}^{d} \ell_i^0(x_i) - \mathrm{a}_0\right)} \mathbf{1}_{\triangle}(x)$$

  such that $\int_{[0,1]} \ell_i^0 q_i \, dx = 0$ with q_i the i-th marginal of the Lebesgue measure on $\triangle$, and a_0 a normalizing constant

- Suppose that for all $1 \leq i \leq d$, ℓ_i^0 belong to a Sobolev space $W_{r_i}^2(q_i)$ with $r_i \in \mathbb{N}^*$ unknown :

$$W_{r_i}^2(q_i) = \left\{ h \in L^2(q_i); h^{(r_i - 1)} \text{ is abs. cont. and } h^{(r_i)} \in L^2(q_i) \right\}.$$

- The product structure of the density suggests a log-additive model to reduce the d-variate problem to d univariate problems

Log-Additive Exponential Series Estimator

Log-additive exponential family

For $\theta = (\theta_{i,k}; 1 \leq i \leq d, 1 \leq k \leq m_i)$:

$$f_\theta(x) = \exp\left(\sum_{i=1}^{d}\sum_{k=1}^{m_i} \theta_{i,k}\varphi_{i,k}(x_i) - \psi(\theta)\right)\mathbf{1}_\triangle(x)$$

- We require a family of functions $(\varphi_{i,k}(x_i); 1 \leq i \leq d, k \in \mathbb{N})$ adapted to $\triangle$ ("orthonormality" w.r.t. the Lebesgue measure on $\triangle$)

Basis functions

For $1 \leq i \leq d$, $k \in \mathbb{N}$, we define for $t \in I$:

$$\varphi_{i,k}(t) = \rho_{i,k} P_k^{(d-i,i-1)}(2t-1),$$

where $P_k^{(d-i,i-1)}$ is the k-th degree Jacobi polynomial and $\rho_{i,k}$ a constant.

Maximum likelihood estimator

- We have a sample of size n : $\mathcal{X}^n = \left(X^j = (X_1^j, \ldots, X_d^j)\right)_{j=1..n}$
- Maximum likelihood estimator $\hat{f}_{m,n} = f_{\hat{\theta}_{\boldsymbol{m},\boldsymbol{n}}}$ verifies, for $1 \leq i \leq d$, $1 \leq k \leq m_i$:

$$\mathbb{E}_{f_{\hat{\theta}_{\boldsymbol{m},\boldsymbol{n}}}}\left[\varphi_{i,k}(X_i)\right] = \hat{\mu}_{m,n,i,k} = \underbrace{\frac{1}{n}\sum_{j=1}^{n}\varphi_{i,k}(X_i^j)}_{\text{empirical mean}}$$

- This is equivalent to (with $|m| = \sum_{i=1}^{d} m_i$) :

$$\begin{aligned}\hat{\theta}_{m,n} &= \operatorname{argmax}_{\theta\in\mathbb{R}^{|\boldsymbol{m}|}} \theta \cdot \hat{\mu}_{m,n} - \psi(\theta) \\ &= \operatorname{argmax}_{\theta\in\mathbb{R}^{|\boldsymbol{m}|}} \underbrace{\frac{1}{n}\sum_{j=1}^{n}\log(f_\theta(X^j))}_{\text{log-likelihood}}\end{aligned}$$

Result of non-adaptive convergence rate I.

Theorem

Let $f^0(x) = \exp\left(\sum_{i=1}^d \ell_i^0(x_i) - a_0\right) \mathbf{1}_\triangle(x)$. *Assume that* $\ell_i^0 \in W_{r_i}^2(q_i)$, $r_i \in \mathbb{N}$, $r_i > d$. *Choose* $m_i = m_i(n) \to \infty$ *such that :*

$$|m|^{2d} \sum_{i=1}^d m_i^{-2r_i} \to 0 \quad \text{and} \quad |m|^{2d+1} / n \to 0,$$

then the Kullback-Leibler divergence between f *and* $f_{\hat{\theta}}$ *satisfies :*

$$D(f^0 || \hat{f}_{m,n}) = O_{\mathbb{P}}\left(\sum_{i=1}^d \left(m_i^{-2r_i} + \frac{m_i}{n}\right)\right)$$

Result of non-adaptive convergence rate II.

Optimal convergence rate

If we choose m_i proportional to $n^{1/(2r_i+1)}$, we obtain the optimal univariate rate :

$$D(f||\hat{f}_{m,n}) = O_{\mathbb{P}}\left(\sum_{i=1}^{d} n^{\frac{-2r_i}{2r_i+1}}\right) = O_{\mathbb{P}}\left(n^{\frac{-2\min(r)}{2\min(r)+1}}\right)$$

Same rate with $m_i = n^{1/(2\min(r)+1)}$ for all $1 \leq i \leq d$

Uniform convergence

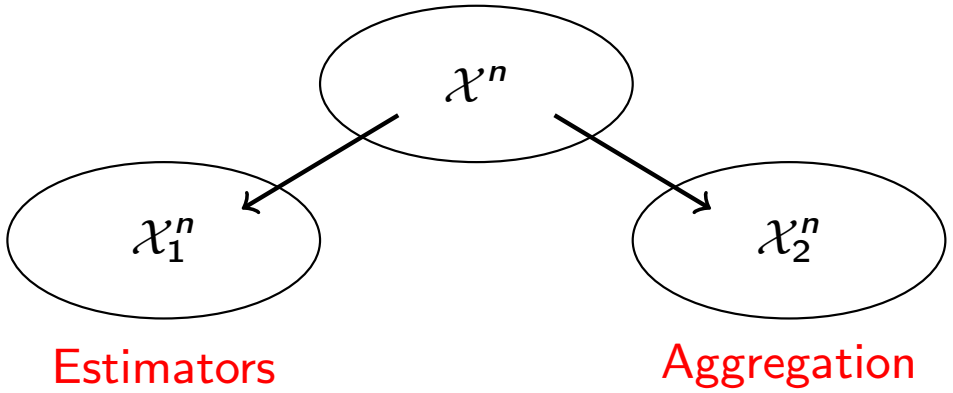
$$\mathcal{K}_r(\kappa) = \left\{ f^0 = \mathrm{e}^{\sum_{i=1}^{d} \ell^0_{[i]} - a_0}; \| \ell_i^0 \|_\infty \leq \kappa, \| (\ell_i^0)^{(r_i)} \|_{L^2(q_i)} \leq \kappa \right\}$$

The convergence in probability is uniform on the set $\mathcal{K}_r(\kappa)$ of densities :

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{f^0 \in \mathcal{K}_r(\kappa)} \mathbb{P}\left(D\left(f^0 \| \hat{f}_{m,n}\right) \geq \left(\sum_{i=1}^{d} m_i^{-2r_i} + \frac{|m|}{n} \right) K \right) = 0$$

Adaptive estimation

- The optimal choice $m_i \sim n^{1/(2\min(r)+1)}$ depends on r, which is unknown
- Adaptation method :
  1. Split the sample into two parts :



  2. Create multiple estimators $\hat{f}_{m,n} = f_{\hat{\theta}_{m,n}}$ with $m \in \mathcal{M}_n$ based on the sample $\mathcal{X}_1^n$
     - Number of estimators : N_n, increasing with n
     - Each $m \in \mathcal{M}_n$ corresponds to regularity parameters r with $\min(r)$ fixed
  3. Perform a convex aggregation on the logarithms of $\hat{f}_{m,n}$ with the sample $\mathcal{X}_2^n$ to obtain the final estimator $f_{\hat{\lambda}_n^*}$

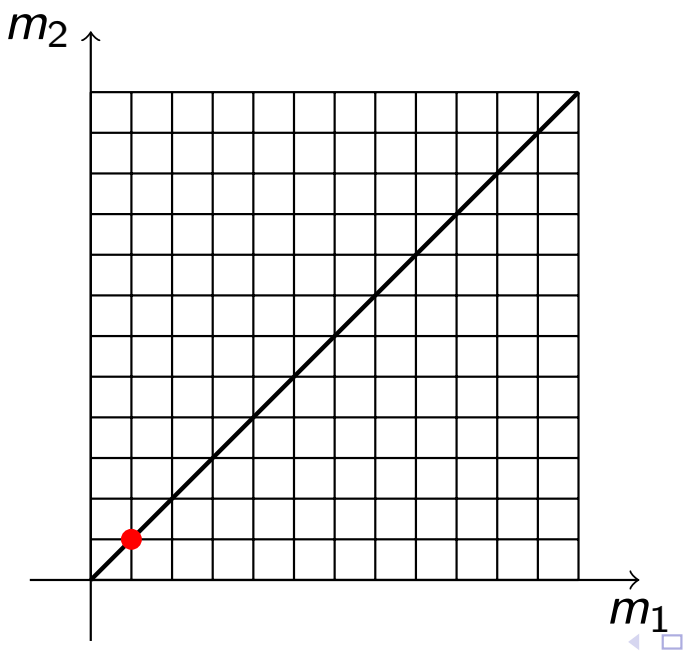
Choice of estimators

- Number of estimators $N_n = o(\log(n))$, $\lim_{n\to\infty} N_n = +\infty$
- The grid :

$$\mathcal{N}_n = \left\{ \lfloor n^{\frac{1}{2(d+j)+1}} \rfloor, 1 \leq j \leq N_n \right\}$$

- Same number of basis functions in each direction :

$$\mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}$$



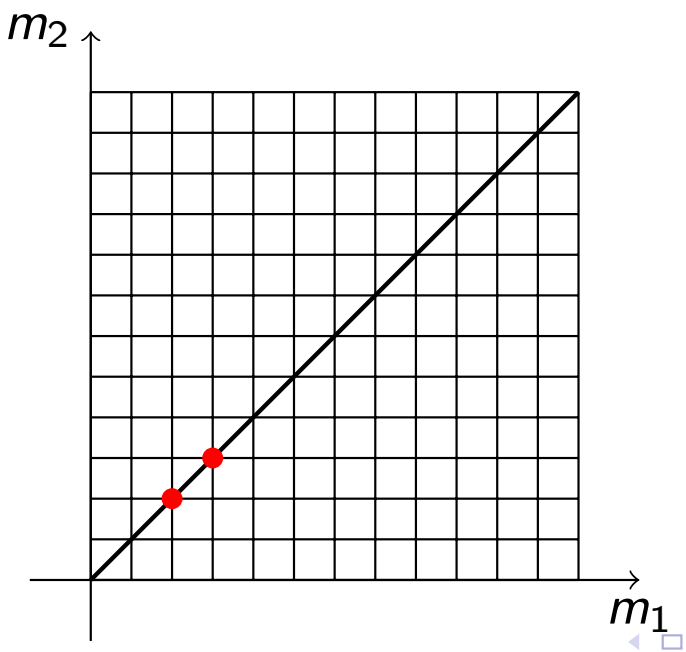
Choice of estimators

- Number of estimators $N_n = o(\log(n))$, $\lim_{n\to\infty} N_n = +\infty$
- The grid :

$$\mathcal{N}_n = \left\{ \lfloor n^{\frac{1}{2(d+j)+1}} \rfloor, 1 \leq j \leq N_n \right\}$$

- Same number of basis functions in each direction :

$$\mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}$$



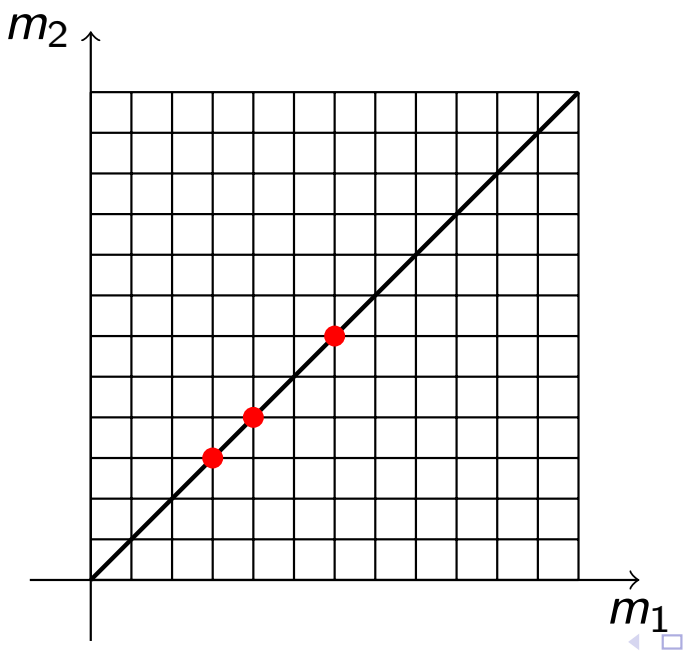
Choice of estimators

- Number of estimators $N_n = o(\log(n))$, $\lim_{n\to\infty} N_n = +\infty$
- The grid :

$$\mathcal{N}_n = \left\{ \lfloor n^{\frac{1}{2(d+j)+1}} \rfloor, 1 \leq j \leq N_n \right\}$$

- Same number of basis functions in each direction :

$$\mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}$$



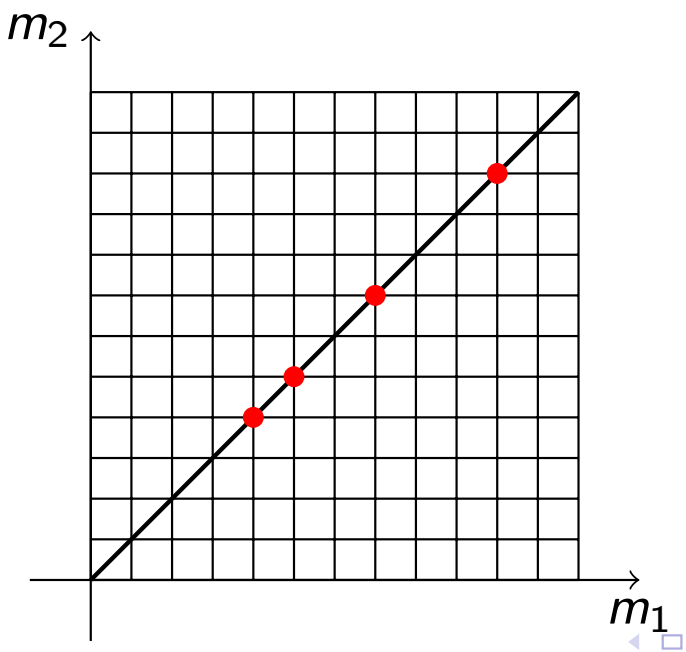
Choice of estimators

- Number of estimators $N_n = o(\log(n))$, $\lim_{n\to\infty} N_n = +\infty$
- The grid :

$$\mathcal{N}_n = \left\{ \lfloor n^{\frac{1}{2(d+j)+1}} \rfloor, 1 \leq j \leq N_n \right\}$$

- Same number of basis functions in each direction :

$$\mathcal{M}_n = \left\{ m = (v, \ldots, v) \in \mathbb{R}^d, v \in \mathcal{N}_n \right\}$$



Convex aggregation of log-densities

Convex combination of log-densities

Let $\hat{\ell}_{m,n}(x) = \sum_{i=1}^{d} \sum_{k=1}^{m_i} \hat{\theta}_{i,k} \varphi_{i,k}(x_i)$ for $m \in \mathcal{M}_n$

$$f_\lambda(x) = \exp\left(\sum_{m \in \mathcal{M}_n} \lambda_m \hat{\ell}_{m,n}(x) - \psi_\lambda \right) \mathbf{1}_\triangle(x)$$

with $\lambda \in \Lambda^+ = \{(\lambda_m, m \in \mathcal{M}_n), \lambda_m \geq 0 \text{ and } \sum_{m \in \mathcal{M}_n} \lambda_m = 1\}$

- Selection of weights $\hat{\lambda}_n^*$ based on the sample $\mathcal{X}_2^n$:

$$\hat{\lambda}_n^* = \underset{\lambda \in \Lambda^+}{\operatorname{argmax}} \underbrace{\frac{1}{|\mathcal{X}_2^n|} \sum_{X^j \in \mathcal{X}_2^n} \log\left(f_\lambda(X^j)\right)}_{\text{log-likelihood}} - \underbrace{\frac{1}{2} \operatorname{pen}(\lambda)}_{\text{penalty}}$$

with $\operatorname{pen}(\lambda) = \sum_{m \in \mathcal{M}_n} \lambda_m D\left(f_\lambda \| \hat{f}_{m,n}\right)$

Sharp oracle inequality for aggregation

Lemma

Let $n \in \mathbb{N}^*$ be fixed. The convex aggregate estimator $f_{\hat{\lambda}_n^*}$ verifies for any $x > 0$ with probability greater than $1 - \exp(-x)$:

$$D\left(f^0 \| f_{\hat{\lambda}_n^*}\right) - \min_{m \in \mathcal{M}_n} D\left(f^0 \| \hat{f}_{m,n}\right) \leq \frac{\beta(\log(N_n) + x)}{n},$$

with a constant $\beta = \beta(\| \ell^0 \|_\infty, \| (\ell_i^0)^{(r_i)} \|_{L^2(q_i)})$.

- Order of the remainder term $\log(N_n)/n$ negligible compared to $n^{-2\min(r)/(2\min(r)+1)}$.

Adaptive estimation - Main result

Theorem

The convex aggregate estimator $f_{\hat{\lambda}_n^*}$ *converges to* f *in probability with the convergence rate :*

$$D(f||f_{\hat{\lambda}_n^*}) = O_{\mathbb{P}}\left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right).$$

Uniform convergence

The convergence is uniform for $r \in \mathcal{R}_n = \{j, d+1 \leq j \leq R_n\}$:

$$\lim_{K\to\infty} \limsup_{n\to\infty} \sup_{r\in(\mathcal{R}_n)^d} \sup_{f^0\in\mathcal{K}_r(\kappa)} \mathbb{P}\left(D\left(f^0\|f_{\hat{\lambda}_n^*}\right) \geq \left(n^{-\frac{2\min(r)}{2\min(r)+1}}\right) K\right) = 0,$$

where R_n satisfies :

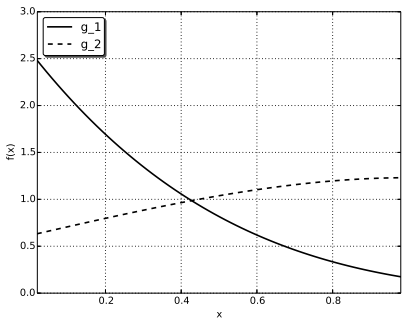
$$R_n \leq N_n + d, \quad R_n \leq \left\lfloor n^{\frac{1}{2(d+N_n)+1}} \right\rfloor, \quad R_n \leq \frac{\log(n)}{2\log(\log(N_n))} - \frac{1}{2}$$

Summary

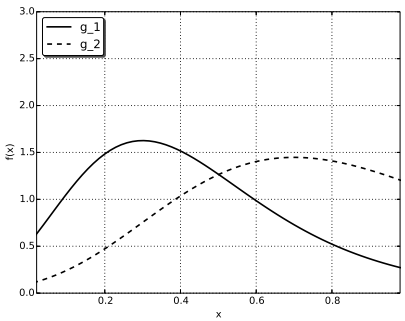
Truncation model

- $Y = (Y_1, Y_2)$ independent with density p_1, p_2
- We observe only when $0 \leq Y_1 \leq Y_2 \leq 1$
- Density of the observations :

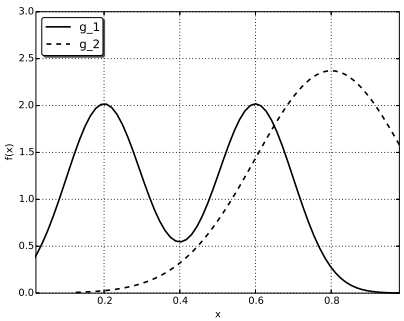
$$f(x) = \frac{p_1(x_1)p_2(x_2)}{\int_{\triangle} p_1(x_1)p_2(x_2)\,dx}\mathbf{1}_{\triangle}(x).$$



(a) Beta



(b) Gumbel



(c) Normal mix

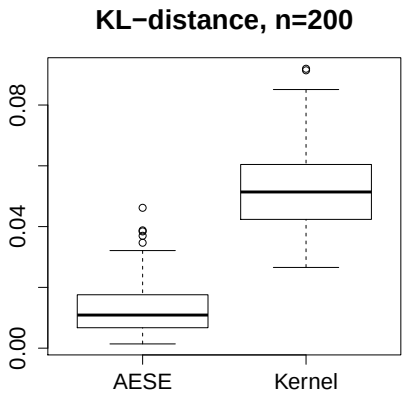
Simulation framework

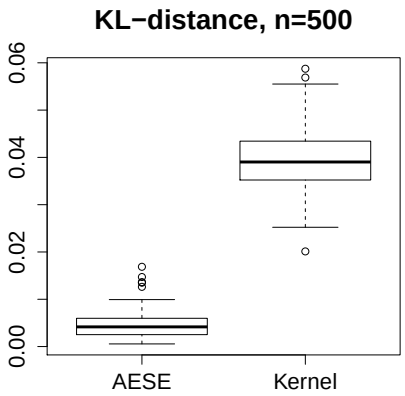
- Aggregate estimator with $m_1 = m_2 = 1, 2, 3, 4$
- Sample size : $n = 200, 500, 1000$
- Split into two parts : $n_1 = 0.8n$, $n_2 = 0.2n$
- Parameter estimation :

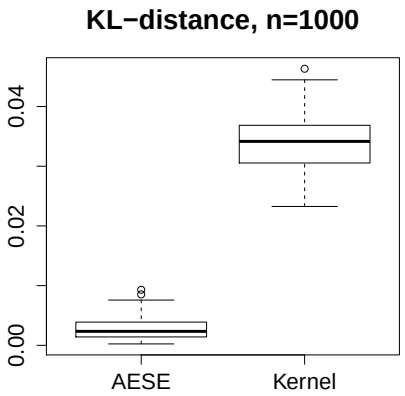
$$\hat{\theta}_{m,n} = \underset{\theta \in \mathbb{R}^{m_1+m_2}}{\operatorname{argmax}} \ \theta \cdot \hat{\mu}_{m,n} - \psi(\theta)$$

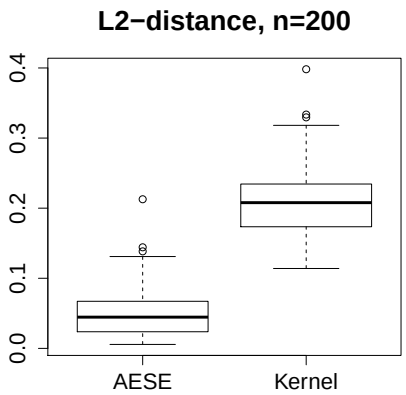
  calculated by numerical optimization
- Comparison with kernel density estimator with Gaussian kernel and bandwidth selected according to Scott's rule
- We calculate the average Kullback-Leibler distance based on 100 estimations

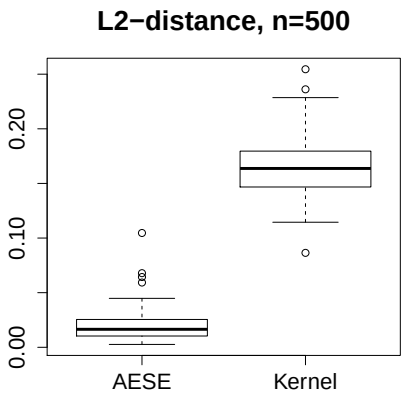
Simulation results - Beta I.

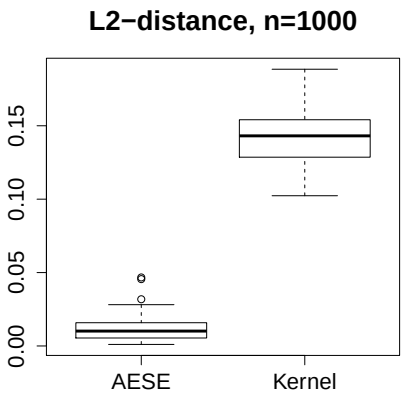




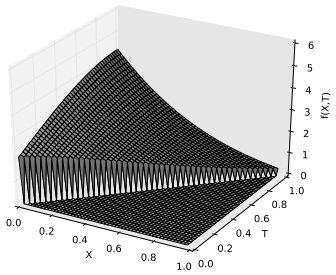




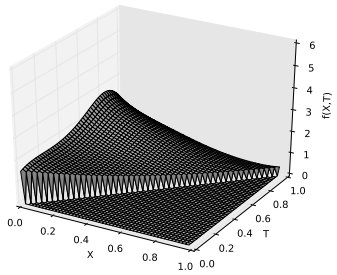




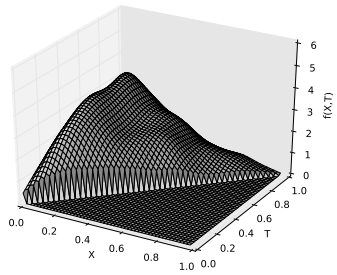
Simulation results - Beta II.



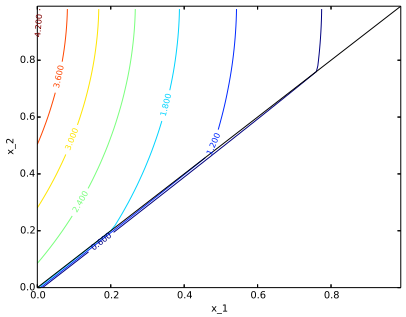
(a) True density

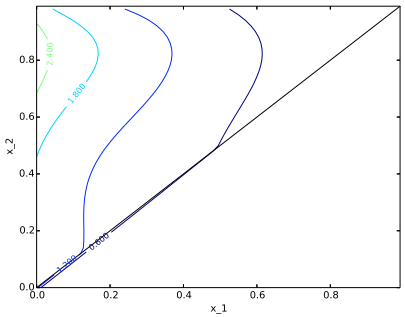


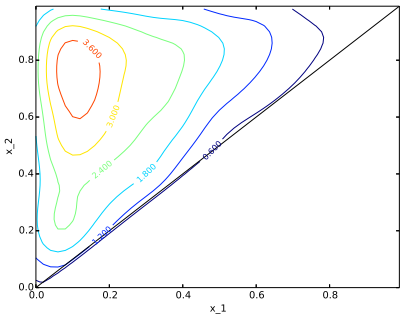
(b) LAESE



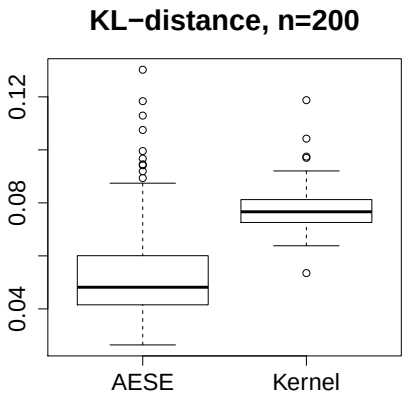
(c) Kernel

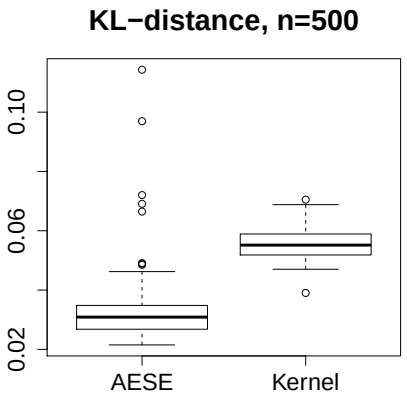


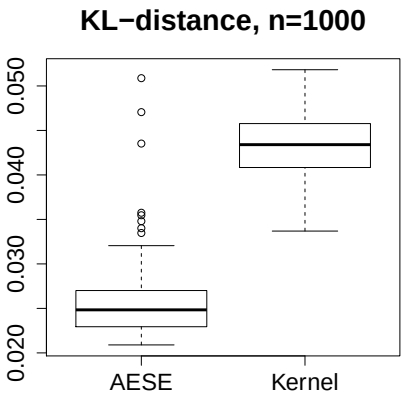


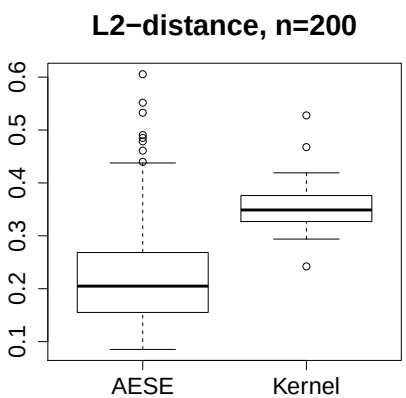


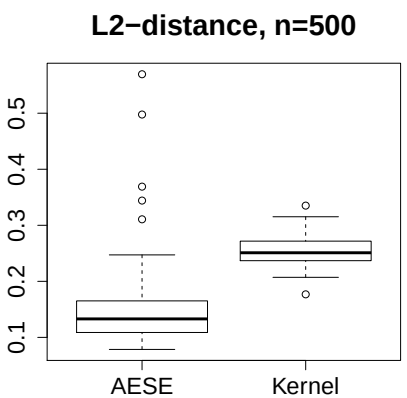
Simulation results - Normal mix I.

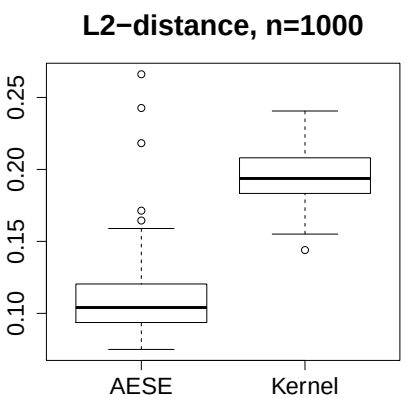




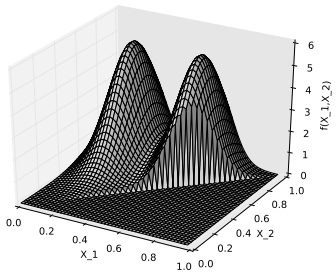




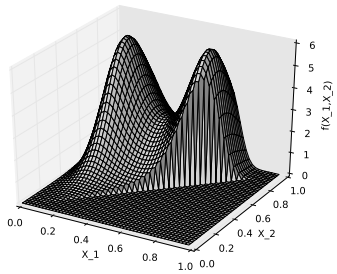




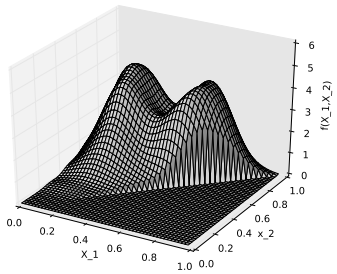
Simulation results - Normal mix II.



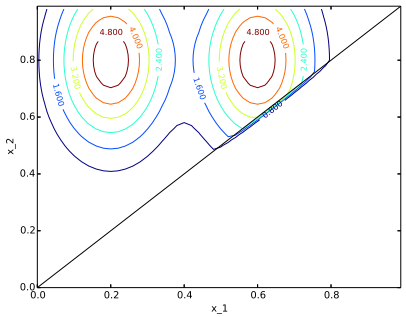
(a) True density

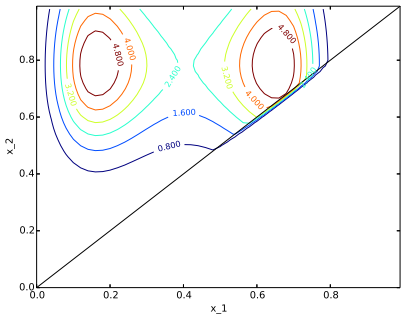


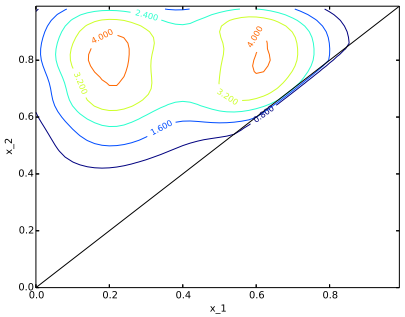
(b) LAESE



(c) Kernel







Bibliography

A.R. Barron,C.-H. Sheu
Approximation of density functions by sequences of exponential families.
The Annals of Statistics, 19(3) :1347–1369, 1991.

Y. Yang, A.R. Barron
Information-theoretic determination of minimax rates of convergence.
Annals of Statistics 1564–1599, 1999.

X. Wu
Exponential series estimator of multivariate densities.
Journal of Econometrics, 156(2) :354–366, 2010.

C. Butucea, J.-F. Delmas, A. Dutfoy, R. Fischer
Nonparametric estimation of distributions of order statistics with application to nuclear engineering
Paper presented at Safety and Reliability of Complex Engineered Systems : ESREL 2015, 2015.

C. Butucea, J.-F. Delmas, A. Dutfoy, R. Fischer
Optimal exponential bounds for aggregation of estimators for the Kullback-Leibler loss
Submitted to Electronic Journal of Statistics, 2016.

C. Butucea, J.-F. Delmas, A. Dutfoy, R. Fischer
Fast adaptive estimation of log-additive exponential models in Kullback-Leibler divergence
Working paper, 2016.