Effective Construction of Modified Histograms in Higher Dimensions

Alain Berlinet, Laurent Rouvière

February 16, 2005

Effective Construction of Modified Histograms in Higher Dimensions

Alain BERLINET and Laurent ROUVIÈRE

Institut de Mathématiques et de Modélisation de Montpellier,
UMR CNRS 5149, Equipe de Probabilités et Statistiques,
Université Montpellier II, Cc 051,
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

Abstract

Density estimation raises delicate problems in higher dimensions especially when strong convergence is required and data marginals can be highly correlated. Modified histograms have been introduced to circumvent the problem of low bin counts when convergence is considered in the sense of information divergence. These estimates are defined from some reference probability density and an associated partition which is defined in the univariate case from the quantiles of the reference density. Therefore, in the multivariate case, the definition of the partition causes an additional problem related to the lack of total order. In this paper, we present a method for constructing modified multivariate histograms such that the corresponding partition is well adapted to the observed data. The approach is based on a data-driven coordinate system selected by cross-validation. We discuss the performance of our estimate with the help of a finite sample simulation study.

1 Introduction

We consider the problem of estimating an unknown probability density f defined on $\mathbb{R}^d$ based on independent, identically distributed observations $X_1, \ldots, X_n$ from f. Here the quality of estimation will be evaluated by a nonnegative divergence $F(f, f_n)$. Of interest are estimators f_n consistent in the sense

$$\lim_{n\to\infty} F(f, f_n) = 0 \; a.s. \quad \text{or} \quad \lim_{n\to\infty} \mathbf{E}F(f, f_n) = 0$$

where $\mathbf{E}$ denotes the expectation with respect to the random vector $(X_1, \ldots, X_n)$ figuring in the estimate f_n. The two most important divergences in

mathematical statistics and information theory are the total variation V and the information divergence D. They are defined by

$$V(f,g) = \frac{1}{2}\int_{\mathbb{R}^d} |f(x) - g(x)|\lambda(dx) = \frac{1}{2}\|f - g\|_{L_1}$$

$$D(f,g) = \begin{cases} \int_{\mathbb{R}^d} f(x) \log \dfrac{f(x)}{g(x)} \lambda(dx) & \text{if } f \ll g \\ \infty & \text{otherwise,} \end{cases}$$

where $\ll$ denotes absolute continuity. It is well known (cf. Csiszàr (1967), Kemperman (1969), and Kullback (1967)) that for all densities f and g, $V(f,g)$ and $D(f,g)$ are linked by the following inequality, called Kullback-Csiszár-Kemperman inequality:

$$2V^2(f,g) \leq D(f,g)\,,$$

which entails that the information divergence is topologically stronger than the total variation. In numerous application fields of statistics (data compression, telecommunication networks, classification, pattern recognition, neural networks...), the consistency defined by total variation may prove inadequate. This is the case when precise estimation of tail probabilities or convergence of integrals of various functionals are required (see Berlinet, Vajda and van der Meulen (1998) for discussion). Another concern with convergence in total variation is that, given any sequence of density estimates, the rate of convergence of the expected L_1 error can be arbitrary slow (Devroye, 1983). Therefore stronger topologies such as information divergence are often preferred.
Classical nonparametric density estimates such as kernel estimates and histograms are not universally consistent in information divergence (see Hall, 1987). The modified histograms introduced by Barron (1988) and Barron, Györfi and van der Meulen (1992) circumvent this problem. They are defined as follows.

Suppose that we observe independent $\mathbb{R}^d$-valued random variables $X_1, \ldots, X_n$ with common unknown density f.

- Denote by g a known density on $\mathbb{R}^d$ and by ν the associated probability measure;
- Define a sequence of integers $\{m_n\}_{n\geq 1}$ such that $1 \leq m_n \leq n$ and let $h_n = 1/m_n$;
- Introduce a sequence of partitions $\mathcal{P}_n = \{A_{n,1}, A_{n,2}, \ldots, A_{n,m_n}\}$, $n \geq 1$, such that $\nu(A_{n,i}) = h_n$, $i = 1, \ldots, m_n$;

- Finally consider, for $a_n = 1/(nh_n + 1)$ the following estimator f_n

$$f_n(x) = \Big[(1 - a_n)\frac{\mu_n\big(A_n(x)\big)}{h_n} + a_n\Big]g(x) = \frac{n\mu_n\big(A_n(x)\big) + 1}{nh_n + 1}g(x)\,. \quad (1.1)$$

  where μ_n stands for the empirical measure associated with the sample $X_1, \ldots, X_n$ and $A_n(x)$ stands for $A_{n,i}$ if $x \in A_{n,i}$.

The estimate (1.1) is a mixture of a histogram-type density estimate and the known density g. It can also be regarded as a piecewise transformation of g itself, which is thus often called in this context the *reference density*.
Under the conditions

$$D(f, g) < \infty, \qquad \lim_{n\to\infty} h_n = 0 \text{ and } \lim_{n\to\infty} nh_n = \infty\,,$$

almost sure consistency in information divergence and consistency in expected information divergence have been proved by Barron, Györfi and van der Meulen (1992).
For further results on modified histograms we refer the reader to Berlinet and Brunel (2004), Berlinet, Györfi and van der Meulen (1997), Berlinet and Biau (2004) and Györfi, Liese, Vajda and van der Meulen (1998).

When $d = 1$, the quantiles of the reference density are used to partition $\mathbb{R}$. Formally, denoting by G the distribution function associated with the probability density g (g is defined on $(a; b)$, a and b may be infinite), we set

$$A_{n,i} = \Big(G^{-1}\Big(\frac{i-1}{m_n}\Big), G^{-1}\Big(\frac{i}{m_n}\Big)\Big],\ i = 1, \ldots, m_n\,,$$

where the interval $(.,.]$ is understood as open on the left and closed on the right only when its upper bound is finite and where G^{-1} is the quantile function defined by

$$\begin{cases} G^{-1}(\alpha) = \inf\{x : G(x) \geq u\} & \text{if } 0 < \alpha < 1 \\ G^{-1}(\alpha) = a & \text{if } \alpha = 0 \\ G^{-1}(\alpha) = b & \text{if } \alpha = 1\,. \end{cases} \quad (1.2)$$

Thus, univariate modified histograms result from the comparison of the quantiles of g with the empirical quantiles. Under mild conditions the choice of g does not affect dramatically the asymptotics. Practically, however, g should not be "too far" from f, so that the comparison between the empirical measure and the reference density over the partition makes sense.

For $d \geq 2$, the choice of such a partition is much more delicate because the lack of total order does not allow to define multivariate quantiles having

the same properties as univariate ones. The aim of this paper is to propose a method for constructing multivariate modified histograms. In Section 2, we give two algorithms to construct this estimate. The first one uses rectangles to partition $\mathbb{R}^d$ (as for the standard multivariate regular histogram estimate). However, the performance of this estimate becomes poor in the presence of high correlation among components of the data vector. This leads us to a more effective method which results from a transformation of these rectangles. We use the data-driven coordinate system introduced by Chaudhuri and Sengupta (1993). In Section 3, we select this coordinate system by cross-validation and we end with some simulations showing the very good performance of the second estimate.

2 Construction of the estimator

Not any sequence of partitions of $\mathbb{R}^d$ has good properties to build consistent estimates. The following concept, introduced by Csizàr (1973) has a great importance in the definition of suitable partitions.

Definition 2.1 *A sequence of partitions $\{\mathcal{P}_n\}$ of $\mathbb{R}^d$ is said to be ν-approximating for a given probability measure ν if, for every measurable set A and for every $\epsilon > 0$, there is for all n sufficiently large a set A_n equal to a union of sets in $\{\mathcal{P}_n\}$ such that*

$$\nu(A_n \triangle A) < \epsilon\,,$$

where $A_n \triangle A$ denotes the symmetric difference of A_n and A.

As proved by Barron, Györfi and van der Meulen (1992) this notion is basic in the proof of consistency of modified histograms.
The partition of a univariate modified histogram is computed from the quantiles of the reference density. Several authors have proposed extensions of quantiles to multidimensional spaces. Chaudhuri (1996) proposed the notion of geometric quantile which generalizes the spatial median studied earlier (see Brown (1983) and Kemperman (1987)). Chakraborty (2001) transformed these geometric quantiles in order to obtain affine equivariant multivariate quantiles. Liu, Parelius and Singh (1999) proposed to define affine equivariant multivariate quantiles using depth analysis. They generalized half-space depth quantiles introduced by Tuckey (1975). Given a measure ν, using *quantile contour plots* of Chakraborty (2001) or *center outward quantiles surfaces* of Liu, Parelius and Singh (1999), one can construct a sequence of partitions $\mathcal{P}_n = \{A_{n,1}, \ldots, A_{n,m_n}\}$ such that $\nu(A_{n,i}) = h_n$ $(i = 1, \ldots, m_n)$. These sequences are nested in the sense that for all n there exists a sequence

$$B_{n,1} \subset B_{n,2} \subset \ldots \subset B_{n,m_n}$$

such that

$$\forall i = 1, \ldots, m_n, \ A_{n,i} = B_{n,i} - \bigcup_{j=1}^{i-1} B_{n,j} \, . \tag{2.1}$$

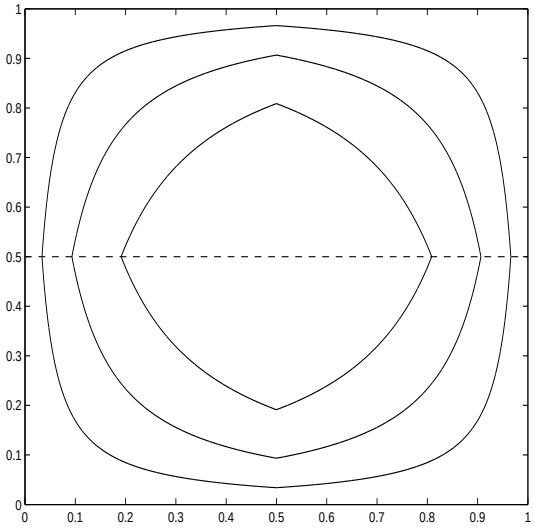

Figure 1: Half-space depth center-outward quantile surface of order 0.25, 0.5 and 0.75 for the uniform distribution on the square $[0,1]^2$.

Such a sequence of partitions is not ν-approximating for any measure ν. For example, let ν be the uniform distribution on the square $[0,1]^2$ and consider a sequence of partitions built from halfspace depth quantiles (see Liu, Parelius and Singh (1999)). Formally, for $i = 1, \ldots, m_n$, $B_{n,i}$ is *a half-space depth center-outward quantile surface* of order i/m_n and $A_{n,i}$ is defined by (2.1) (see Figure 1). Consider the vertical line which passes through the center of the square (dashed line in Figure 1). This line splits the square into two rectangles. If A denotes one of these rectangles, it is easily seen that for all sets A_n equal to a union of sets in $\mathcal{P}_n$, we have

$$\nu(A_n \, \triangle \, A) = 0.5 \, ,$$

which entails that $\mathcal{P}_n$ is not ν-approximating.
Other authors have defined quantiles in multidimensional spaces (see Brown and Hettmansperger (1987), Eddy (1983, 1985)), but as far as we know none permits the construction of modified histograms for any reference density. This leads us to restrict our attention to a certain class of reference densities.

2.1 Regular modified histograms

The standard regular (unmodified) histogram is defined by a partition of $\mathbb{R}^d$ into rectangular cells of widths $h_1, \ldots, h_d$. The goal of this paragraph is the adaptation of this partition to modified histograms. In this regard we only consider reference densities g such that

$$g(x_1, \ldots, x_d) = g_1(x_1) \ldots g_d(x_d)\,, \tag{2.2}$$

where $g_1, \ldots, g_d$ are univariate densities. For $j = 1, \ldots, d$, we denote by G_j the distribution function associated with the probability density g_j and by G_j^{-1} the quantile function as in (1.2).
Given i.i.d. observations $X_1, \ldots, X_n$ from a density f on $\mathbb{R}^d$ and given a reference density g such as (2.2), modified multivariate histograms are built as follows:

- Set $m = m_1 \ldots m_d$ with $m_1, \ldots, m_d$ positive integers and let $h_j = 1/m_j$ for $j = 1, \ldots, d$;
- For $j = 1, \ldots, d$ and $i_j = 1, \ldots, m_j - 1$, compute univariate quantiles of order $i_j h_j$ of g_j. Denote by q_{j,i_j} these quantiles *i.e.*

  $$q_{j,i_j} = G_j^{-1}(i_j h_j)$$

  with the convention $q_{j,0} = -\infty$ and $q_{j,m_j} = \infty$;
- Consider the grid defined by the above family $\{q_{j,i_j}\}$; this grid leads to a partition of $\mathbb{R}^d$ into m hyperrectangles (see Figure 2), say

  $$A_{i_1,\ldots,i_d} = \prod_{j=1}^{d} (q_{j,i_j-1}, q_{j,i_j}]\,; \tag{2.3}$$

- For each of these cells, compute the empirical measure:

  $$\mu_n(A_{i_1,\ldots,i_d}) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \in A_{i_1,\ldots,i_d}\}}\,;$$

- The *regular modified multivariate histogram density estimate* f_n is defined by:

  $$f_n(x) = \frac{n\mu_n\big(A(x)\big) + 1}{nh + 1} g(x) \tag{2.4}$$

  where $h = h_1 \ldots h_d$ and $A(x)$ stands for $A_{i_1,\ldots,i_d}$ if $x \in A_{i_1,\ldots,i_d}$.

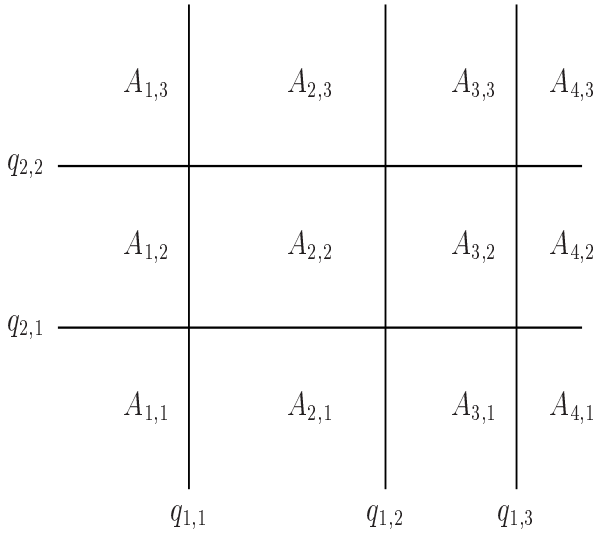

Figure 2: Example of partition in $\mathbb{R}^2$: $m_1 = 4$, $m_2 = 3$.

Denote by ν the probability measure associated with the reference density g. It is easily seen that, for any set $A_{i_1,\ldots,i_d}$,

$$\nu(A_{i_1,\ldots,i_d}) = h\,.$$

Consistency in information divergence and expected information divergence is established in our next theorem.

Theorem 2.1 *Let f_n be the regular modified histogram defined in* (2.4)*. Assume that $D(f,g) < \infty$.*

(i) If $h_j = h_{j,n}$ $(j = 1,\ldots,d)$ and $\lim_{n\to\infty} \max_{1\le j\le d} h_{j,n} = 0$ then the sequence of partition

$$\{\mathcal{P}_n\} = \{A_{n,i_1,\ldots,i_d}\}_{\substack{1\le j\le d \\ 1\le i_j\le m_{j,n}}}$$

defined in (2.3) *is ν-approximating.*

(ii) Moreover assume that $\lim_{n\to\infty} nh_n = \infty$ $(h_n = h_{1,n}\ldots h_{d,n})$, then

$$\lim_{n\to\infty} \mathbf{E}D(f,f_n) = 0 \text{ and } \lim_{n\to\infty} D(f,f_n) = 0 \text{ a.s.}$$

Proof .
We first prove (i). Let S denote the support of ν and $\overline{S}$ its complement in $\mathbb{R}^d$. With a slight abuse of notation, we denote

$$\{\mathcal{P}_n\} = \{A_{n,1}, \ldots, A_{n,m_n}\}\,.$$

For $j = 1, \ldots, d$, let $a_j = \inf\{x \in \mathbb{R} : g_j(x) \neq 0\}$ and $b_j = \sup\{x \in \mathbb{R} : g_j(x) \neq 0\}$ (a_j and b_j may be infinite). Let S_j (resp. $\overline{S}_j$) be the projection of S (resp. $\overline{S}$) on (a_j, b_j). $\overline{S}_j$ is the union of k_j distinct intervals of length $l(i)$ $(i = 1, \ldots, k_j)$. For $x = (x_1, \ldots, x_d) \in S$, let $p_j(x_j)$ denote the number of intervals of $\overline{S}_j$ before x_j and consider for $j = 1, \ldots, d$

$$\begin{aligned} T_j : \quad S_j &\longrightarrow \mathbb{R} \\ x_j &\longrightarrow x_j - \sum_{i=1}^{p_j(x_j)} l(i) \end{aligned}$$

and

$$\begin{aligned} T : \qquad S &\longrightarrow \mathbb{R}^d \\ (x_1, \ldots, x_d) &\longrightarrow (T_1(x_1), \ldots, T_d(x_d))\,. \end{aligned}$$

The application T allows to remove the hyperrectangles R of $\mathbb{R}^d$ such that $\nu(R) = 0$.
Fix a measurable set A. If A_n is equal to a union of sets in $\mathcal{P}_n$ then

$$\nu(A_n \vartriangle A) = \nu(T(A_n) \vartriangle T(A))\,.$$

Therefore, it suffices to prove that the partition

$$\{\mathcal{P}_n^T\} = \{T(A_{n,1}), \ldots, T(A_{n,m_n})\}$$

is ν-approximating. Note that $T(A_{n,i})$ $(i = 1, \ldots, m_n)$ are hyperrectangles of $\mathbb{R}^d$ such that $\nu(T(A_{n,i})) = h_{1,n} \ldots h_{d,n}$. Since $\lim_{n\to\infty} h_{j,n} = 0$, $j = 1, \ldots, d$, we have for each ball B centered at some point x_0

$$\lim_{n\to\infty} \max_{\{i:T(A_{n,i})\cap B\neq\emptyset\}} \operatorname{diam}(T(A_{n,i})) = 0$$

where $\operatorname{diam}(E) = \sup_{x,y\in E} d(x, y)$ and $d(x, y)$ denotes the distance in $\mathbb{R}^d$. It follows from Csiszár (1973, p. 168) that the partition $\{\mathcal{P}_n^T\}$ is ν-approximating. Combining (i) with Theorem 2 in Barron, Györfi and van der Meulen (1992) gives (ii).

■

Table 1: Information divergence and total variation according to the correlation.

ρ	0	0.25	0.5	0.75	0.95
$D(f, f_n)$	0.32	0.33	0.35	0.43	0.74
$V(f, f_n)$	0.19	0.19	0.21	0.22	0.34

2.2 Influence of correlation

Through an example, we study the influence of the shape of the data vector on the performance of the density estimate defined in (2.4). Table 1 gives the information divergence $D(f, f_n)$ and the total variation $V(f, f_n)$ between binormals and their standard modified histogram estimates. Simulated binormals have 0 mean, unit standard deviation and varying correlation (from 0 to 0.95), the size of the samples is $n = 250$. To construct the estimate, we take $m_1 = m_2 = 5$ and the reference density g is a product of Gumbel densities:

$$g(x, y) = \exp\big(-x - \exp(-x)\big) \exp\big(-y - \exp(-y)\big). \tag{2.5}$$

Results are clearly better in the presence of weak correlation. One can explain it as follows. On Figure 3, we have represented a sample of size $n = 250$ from a binormal with 0 mean and identity variance matrix (LEFT) and the image of this sample by the affine transformation (RIGHT):

$$T(x) = \Sigma^{1/2} x + a$$

where

$$\Sigma = \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix} \text{ and } a = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

Note that the transformed sample can be seen as a sample simulated from a binormal $\mathcal{N}(a, \Sigma)$. We represent on these graphics the partition used to construct regular modified histograms with a reference density of Gumbel (see (2.5)) and $m_1 = m_2 = 5$. For the transformed sample, only few classes possesses observations, the partition is not well adapted to the data cloud. Therefore the comparison between the empirical measure and the reference density over the partition does not make much sense.
To correct this, we will construct data dependent modified histograms for which keeping the parameters g and m_j $(j = 1, \ldots, d)$ fixed, the corresponding partition is equivariant under affine transformation of data vectors. Our method is inspired by the affine equivariant *quantile contour plots* defined by Chakraborty (2001).

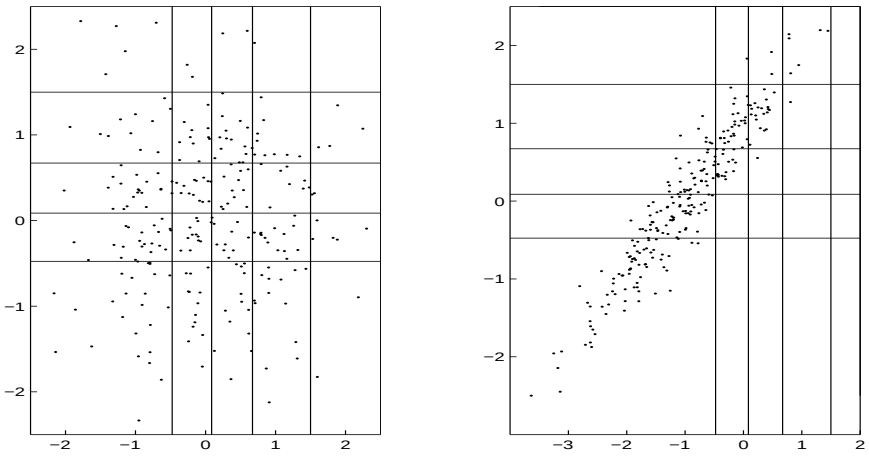

Figure 3: See text in Subsection 2.2.

2.3 Data-driven modified histograms

Statistical practice suggests that histograms based on data-dependent partitions will provide better performance than those based on a fixed sequence of partitions. Theoretical evidence for this superiority was put forward by Stone (1985). In this paragraph, we construct a modified histogram based on a data-dependent partition equivariant under affine transformation of the data vector (for g and m_j, $j = 1, \ldots, d$ fixed). The approach is based on a "data-driven coordinate system" introduced by Chaudhuri and Sengupta (1993). Formally, fix n_0 such that $n_0 > d + 1$ and consider $n_0 + n$ data points $X_1, \ldots, X_{n_0+n}$ i.i.d. from a density f on $\mathbb{R}^d$. Split the data into a set $X_1, \ldots, X_{n_0}$ used for choosing the "data-driven coordinate system" and a set $X_{n_0+1}, \ldots, X_{n_0+n}$ used for constructing the density estimate. To lighten the notation we will write $X_1^\star, \ldots, X_n^\star$ for $X_{n_0+1}, \ldots, X_{n_0+n}$.

- Set $m = m_1 \ldots m_d$ with $m_1, \ldots, m_d$ positive integers, and let $h_j = 1/m_j$ for $j = 1, \ldots, d$;
- Let $\alpha = \{k_0, k_1, \ldots, k_d\}$ denote a subset of $\{1, 2, \ldots, n_0\}$ of size $(d + 1)$. Consider the points $X_{k_0}, \ldots, X_{k_d}$ which will form a "data-driven coordinate system", where X_{k_0} will determine the origin and the lines joining that origin to the remaining d data points $X_{k_1}, \ldots, X_{k_d}$ will form various coordinate axis. Consider the $d \times d$ matrix

$$X(\alpha) = \{X_{k_1} - X_{k_0}, \ldots, X_{k_d} - X_{k_0}\}.$$

  If f is absolutely continuous on $\mathbb{R}^d$, $X(\alpha)$ is an invertible matrix with probability one for any choice of α (see Chaudhuri and Sengupta (1993)).

Next, transform all the observations in terms of the new coordinate system as

$$\begin{cases} \dot{X}_i = \{X(\alpha)\}^{-1} X_i, & i = 1, \ldots, n_0\,, \\ \\ \dot{X}_i^\star = \{X(\alpha)\}^{-1} X_i^\star, & i = 1, \ldots, n\,. \end{cases}$$

- Let $\tilde{g}$ be a density on $\mathbb{R}^d$ such that

$$\tilde{g}(x_1, \ldots, x_d) = \tilde{g}_1(x_1) \ldots \tilde{g}_d(x_d)\,, \tag{2.6}$$

where $\tilde{g}_1(x_1), \ldots, \tilde{g}_d(x_d)$ are univariate densities.
Define $p = (p_1, \ldots, p_d)$ the coordinatewise median associated with the density $\tilde{g}$, *i.e.*

$$p_j = \tilde{G}_j^{-1}(0.5)\,, \quad j = 1, \ldots, d\,,$$

and let $\dot{X}_{([n_0/2])}$ be the empirical coordinatewise median from the sample $\dot{X}_1, \ldots, \dot{X}_{n_0}$ *i.e.*

$$\dot{X}_{([n_0/2])} = (\dot{X}^{(1)}_{([n_0/2])}, \ldots, \dot{X}^{(d)}_{([n_0/2])})\,.$$

where $[\ \ ]$ stands for the integer part and $\dot{X}^{(j)}_{(1)}, \ldots, \dot{X}^{(j)}_{(n_0)}$ denotes the order statistics of the j-th components of the data vector $\dot{X}_1, \ldots, \dot{X}_{n_0}$. Consider the vector $b_X^\alpha = p - \dot{X}_{([n_0/2])}$ and let $\tilde{X}_i^\star$ be the image of $\dot{X}_i^\star$ by the translation of vector b_X^α, *i.e.*

$$\tilde{X}_i^\star = \dot{X}_i^\star + b_X^\alpha\,, \quad i = 1, \ldots, n\,.$$

As for the regular modified histograms presented above, for $j = 1, \ldots, d$ and $i_j = 1, \ldots, m_j - 1$, denote $\tilde{q}_{j,i_j}$ the quantile of order $i_j h_j$ of $\tilde{g}_j$. These quantiles lead to a partition of $\mathbb{R}^d$ into m hyperrectangles say

$$\tilde{A}_{i_1,\ldots,i_d} = \prod_{j=1}^{d} (\tilde{q}_{j,i_j-1}, \tilde{q}_{j,i_j}]\,.$$

Let μ_n (resp. $\tilde{\mu}_n$) be the empirical measure associated with the sample $X_1^\star, \ldots, X_n^\star$ (resp. $\tilde{X}_1^\star, \ldots, \tilde{X}_n^\star$);

- Express the $\tilde{A}_{i_1,\ldots,i_d}$'s in terms of the original coordinate system, *i.e.*

$$A_{i_1,\ldots,i_d} = X(\alpha)\big(\tilde{A}_{i_1,\ldots,i_d} - b_X^\alpha\big)\,.$$

$\tilde{A}_{i_1,\ldots,i_d}$ is the image of the hyperrectangle $A_{i_1,\ldots,i_d}$ by an affine transformation therefore $\tilde{A}_{i_1,\ldots,i_d}$ is an hyperparallelogram (see Figure 4). Moreover, it is easily seen that

$$\mu_n(A_{i_1,\ldots,i_d}) = \tilde{\mu}_n(\tilde{A}_{i_1,\ldots,i_d})\,;$$

- Finally, fix

$$g_\alpha(x) = \frac{1}{\big|det\big(X(\alpha)\big)\big|}\, \tilde{g}\big(\{X(\alpha)\}^{-1}x + b_X^\alpha\big)\,, \tag{2.7}$$

then the *data-driven modified histogram density estimate* is defined by

$$f_n(x) = \frac{n\mu_n\big(A(x)\big) + 1}{nh + 1}\, g_\alpha(x)\,, \tag{2.8}$$

where $h = h_1 \dots h_d$ and $A(x)$ stands for $A_{i_1,\dots,i_d}$ if $x \in A_{i_1,\dots,i_d}$.

Lemma 2.1 *The estimate $f_n(x)$ defined in* (2.8) *is a modified histogram in the sense of* (1.1).

Proof .
It suffices to prove the following assertions:

- g_α is a density (we will denote by ν the measure associated with this density);
- for $j = 1, \dots, d$ and $i_j = 1, \dots, m_j$, $\nu(A_{i_1,\dots,i_d}) = h$.

These assertions are direct consequences of the change of variables theorem. ■

Remark 2.1 One can use other translations, however our choice of b_X^α seems to be well adapted to our estimate. Indeed, modified histograms result from the comparison between the reference density and the empirical measure. Thus, our translation is chosen so that the image of $\dot{X}_1, \dots, \dot{X}_{n_0}$ has the same median as the density $\tilde{g}$. This translation can be seen as a "bias correction". We choose the median because of its robustness.

¿From now on, given a sample $T_1, \dots, T_{n_0+n}$, we write $T_1^\star, \dots, T_n^\star$ for $T_{n_0+1}, \dots, T_{n_0+n}$ and $\mu_n(A; T_1^\star, \dots, T_n^\star)$ for the empirical measure associated with $T_1^\star, \dots, T_n^\star$. Moreover, with a slight abuse of notation, we will denote by $\{A_{i_1,\dots,i_d}\}$ the partition

$$\{A_{i_1,\dots,i_d}\}_{\substack{1\le j\le d \\ 1\le i_j\le m_j}}\,.$$

We now prove the equivariance of the partition under arbitrary affine transformations of data vectors.

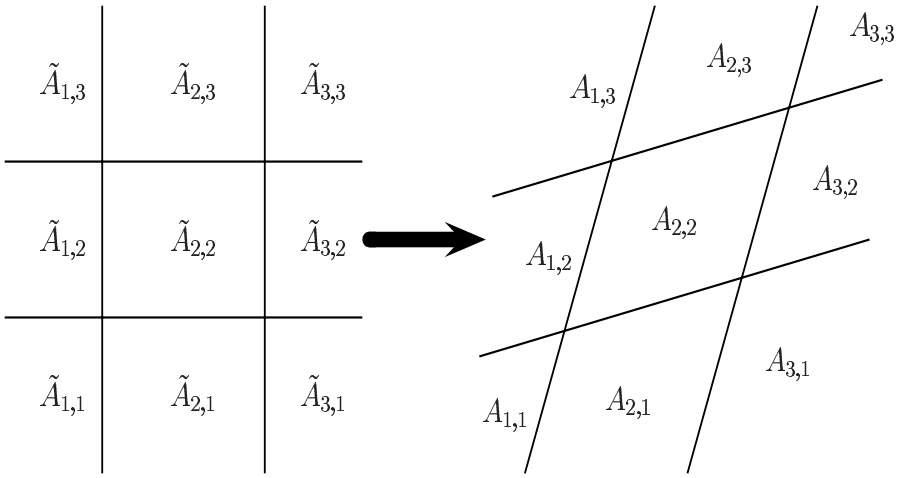

Figure 4: Transformation of the partition.

Theorem 2.2 *The partition* $\{A_{i_1,\ldots,i_d}\}$ *is equivariant under arbitrary affine transformations of data vectors. We can formulate it as follows.*
Let the d-dimensional vectors $X_1, \ldots, X_{n_0+n}$ *be transformed into* $Z_1, \ldots, Z_{n_0+n}$ *with* $Z_i = MX_i + c$ *where* M *is a* $d \times d$ *nonsingular matrix and* c *is a vector in* $\mathbb{R}^d$*. Suppose that we use the same density* $\tilde{g}$ *and the same integers* m_j $(j = 1, \ldots, d)$ *to construct the data-driven modified histogram from* $X_1, \ldots, X_{n_0+n}$ *and* $Z_1, \ldots, Z_{n_0+n}$*. If* $\{A_{i_1,\ldots,i_d}\}$ *(resp.* $\{B_{i_1,\ldots,i_d}\}$*) denotes the partition computed from the sample* $X_1, \ldots, X_{n_0+n}$ *(resp.* $Z_1, \ldots, Z_{n_0+n}$*), then for all integers* $i_1, \ldots, i_d$ *such that* $1 \le i_j \le m_j$ *and* $1 \le j \le d$ *we have*

(i) $B_{i_1,\ldots,i_d} = MA_{i_1,\ldots,i_d} + c\,.$

(ii) $\mu_n(A_{i_1,\ldots,i_d}; X_1^\star, \ldots, X_n^\star) = \mu_n(B_{i_1,\ldots,i_d}; Z_1^\star, \ldots, Z_n^\star)\,.$

Proof .
Let $\alpha = \{k_0, k_1, \ldots, k_d\}$ be a subset of $\{1, \ldots, n_0\}$ of size $d+1$. Consider

$$X(\alpha) = \{X_{k_1} - X_{k_0}, \ldots, X_{k_d} - X_{k_0}\}$$

and

$$Z(\alpha) = \{Z_{k_1} - Z_{k_0}, \ldots, Z_{k_d} - Z_{k_0}\}$$

so that we have $Z(\alpha) = MX(\alpha)$.
Note that for $i = 1, \ldots, n_0$

$$\begin{aligned}\dot{Z}_i =& Z(\alpha)^{-1} Z_i \\ =& X(\alpha)^{-1} X_i + \big(MX(\alpha)\big)^{-1} c \\ =& \dot{X}_i + \big(MX(\alpha)\big)^{-1} c\,.\end{aligned}$$

Therefore $\dot{Z}_{[n_0/2]} = \dot{X}_{[n_0/2]} + (MX(\alpha))^{-1} c$ and $b_Z^\alpha = b_X^\alpha - \big(MX(\alpha)\big)^{-1} c$. As we use the same density $\tilde{g}$ and the same integers m_j $(j = 1, \ldots, d)$, the partitions computed for transformed observations will be the same for the samples $X_1, \ldots, X_{n_0+n}$ and $Z_1, \ldots, Z_{n_0+n}$. We denote $\{\tilde{A}_{i_1,\ldots,i_d}\}$ this partition. To obtain $\{A_{i_1,\ldots,i_d}\}$ and $\{B_{i_1,\ldots,i_d}\}$ we only have to retransform $\{\tilde{A}_{i_1,\ldots,i_d}\}$. For all integers $i_1, \ldots, i_d$ such that $1 \le i_j \le m_j$ and $1 \le j \le d$, it follows that

$$\begin{aligned}B_{i_1,\ldots,i_d} =& Z(\alpha)(\tilde{A}_{i_1,\ldots,i_d} - b_Z^\alpha) \\ =& MX(\alpha)\Big(\tilde{A}_{i_1,\ldots,i_d} - \big(b_X^\alpha - (MX(\alpha))^{-1} c\underline{i}g\big)\Big) \\ =& M\Big(X(\alpha)\big(\tilde{A}_{i_1,\ldots,i_d} - b_X^\alpha\big)\Big) + c \\ =& MA_{i_1,\ldots,i_d} + c\,,\end{aligned}$$

which gives *(i)*.
Since $\tilde{X}_i^\star = X(\alpha)^{-1} X_i^\star + b_X^\alpha$ and $\tilde{Z}_i^\star = Z(\alpha)^{-1} Z_i^\star + b_Z^\alpha$ $(i = 1, \ldots, n)$, it easily follows that

$$\forall i = 1, \ldots, n,\ \tilde{X}_i^\star = \tilde{Z}_i^\star\,.$$

Therefore

$$\mu_n(\tilde{A}_{i_1,\ldots,i_d}; \tilde{X}_1^\star, \ldots, \tilde{X}_n^\star) = \mu_n(\tilde{A}_{i_1,\ldots,i_d}; \tilde{Z}_1^\star, \ldots, \tilde{Z}_n^\star)$$

and *(ii)* is proved.

■

Remark 2.2 It is worth pointing out that the actual reference density (in the sense of (1.1)) is g_α which implies that the reference density depends on the data. However, one can have some a priori idea on the density to estimate and thus want to construct modified histogram for a given reference density g. It is possible to use this algorithm provided that g may be written in the form

$$g(x) = \frac{1}{|\det(M)|}\tilde{g}(M^{-1}x + a) \tag{2.9}$$

where $x = (x_1, \ldots, x_d)$, M is an invertible matrix $d \times d$, a is a vector of $\mathbb{R}^d$ and $\tilde{g}$ is a product of univariate densities *i.e.*

$$\tilde{g}(x) = \tilde{g}_1(x_1) \ldots \tilde{g}_d(x_d)\,.$$

In that case we no longer split the data (all the observations are used to construct the modified histogram) and we replace $X(\alpha)$ by M and b_X^α by a. Note that multinormal densities $g_{\mu,\Sigma}$ are in the form of (2.9). Nevertheless Theorem 2.2 does not hold for such modified histograms.

Summarizing, we have found a partition equivariant under arbitrary affine transformation. Consistency in information divergence of this new estimate is a straightforward consequence of the next lemma (whom proof is straightforward).

Lemma 2.2 *The information divergence is invariant under invertible transformation of the data sample.*

Corollary 2.1 *Let f_n be the data-driven modified histogram defined in* (2.8)*. Assume that $D(f, g_\alpha) < \infty a.s.$ Moreover, assume that for $i = 1, \ldots, d$, $h_i = h_{i,n}$ (therefore $h = h_n$),*

$$\lim_{n\to\infty} \max_{1\le i\le d} h_{i,n} = 0 \text{ and } \lim_{n\to\infty} nh_n = \infty\,,$$

then

$$\lim_{n\to\infty} \mathbf{E}^{(n_0)} D(f, f_n) = 0 a.s. \text{ and } \lim_{n\to\infty} D(f, f_n) = 0 \ a.s.$$

where $\mathbf{E}^{(n_0)}$ denotes the conditional expectation given the X_i's for which $1 \le i \le n_0$.

Proof .
Fix $X_1, \ldots, X_{n_0}$ such that $D(f, g_\alpha) < \infty$. Let $\tilde{f}$ (resp. $\tilde{f}_n$) be the density to estimate f (resp. the density estimator f_n) in the transformed coordinate system *i.e.*

$$\begin{cases} \tilde{f}(x) = & \big|det\big(X(\alpha)\big)\big| f\big(X(\alpha)(x - b_X^\alpha)\big) \\ \\ \tilde{f}_n(x) = & \big|det\big(X(\alpha)\big)\big| f_n\big(X(\alpha)(x - b_X^\alpha)\big)\,. \end{cases}$$

$\tilde{f}_n$ is the regular modified histogram density estimate of $\tilde{f}$ (see page 6) with $\tilde{g}$ as reference density. From Theorem 2.1, it follows that

$$\begin{cases} \lim\limits_{n\to\infty} D(\tilde{f}, \tilde{f}_n) = 0 \ a.s. \\ \\ \lim\limits_{n\to\infty} \mathbf{E} D(\tilde{f}, \tilde{f}_n) = 0\,. \end{cases}$$

The conclusion follows from Lemma 2.2.

■

3 Selection of α

The performance of the data-driven modified histogram clearly depends upon the choice of m_j $(j = 1, \ldots, d)$, $\tilde{g}$ and α. Here we will restrict our attention to the choice of α. Recent univariate results obtained by Berlinet and Brunel (2004) show that the Kullback-Leibler cross-validation technique works well for selecting m_1 from the data. We will apply the same method to find the best α.

Let $X_1, \ldots, X_{n_0+n}$ be i.i.d. observations from a density f and let S_{n_0} denote the collection of all subsets of size $d+1$ of $\{1, \ldots, n_0\}$. Fix m_j $(j = 1, \ldots, d)$ and $\tilde{g}$ (such as (2.6)). For $\alpha \in S_{n_0}$, let us denote by f_n^α the data-driven modified multivariate histogram defined in (2.8). Expanding the actual information divergence error yields

$$D(f, f_n^\alpha) = \int_{\mathbb{R}^d} f(x) \log f(x)\, dx - \int_{\mathbb{R}^d} f(x) \log f_n^\alpha(x)\, dx\,. \tag{3.1}$$

The second integral could be written as $\mathbf{E}(\log f_n^\alpha(X))$, where the expectation is taken with respect to the evaluating point and not over the sample. The cross-validation device consists in removing one data point among $X_1^\star, \ldots, X_n^\star$ and using the remaining $(n-1)$ points to construct an estimator of $\mathbf{E}(\log f_n^\alpha(X))$. This step is repeated for each $X_i^\star$ $(i = 1, \ldots, n)$. Let $f_n^{\alpha,i}$ be the modified histogram density estimate defined after deleting the i-th observation *i.e.*

$$f_n^{\alpha,i}(x) = \frac{n\mu_n^i\big(A(x)\big) + 1}{nh + 1} g_\alpha(x)$$

where g_α is defined by (2.7) and

$$\mu_n^i\big(A(x)\big) = \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}_{\{X_j^\star \in A(x)\}}\,.$$

With this notation, an estimate of $\mathbf{E}(\log f_n^\alpha(X))$ is given by

$$\frac{1}{n} \sum_{i=1}^{n} \log f_n^{\alpha,i}(X_i^\star)$$

and since the first integral in (3.1) does not depend on α, we deduce a cross validation criterion for the choice of α:

choose $\hat{\alpha} \in S_{n_0}$ which minimizes $CV(\alpha) = -\frac{1}{n}\sum_{i=1}^{n} \log\, f_n^{\alpha,i}(X_i^\star)$.

Note that if $D(f, g_{\hat{\alpha}}) < \infty$ *a.s.*, consistency of the selected estimate $f_n^{\hat{\alpha}}$ follows from Corollary 2.1.
For fixed α, we have seen that the partition is affine equivariant. The next theorem states the analogue with α selected by cross-validation.

Theorem 3.1 *The choice of α by cross-validation is invariant under arbitrary affine transformations of data vectors. We can formulate it as follows. Let the d-dimensional vectors $X_1, \ldots, X_{n_0+n}$ be transformed into $Z_1, \ldots, Z_{n_0+n}$ with $Z_i = MX_i + c$ where M is a $d \times d$ nonsingular matrix and c is a vector in $\mathbb{R}^d$. Suppose that we use the same density $\tilde{g}$ and the same integers m_j $(j = 1, \ldots, d)$ to construct the data-driven modified histograms $f_{n,X}^{\alpha}$ (with $X_1, \ldots, X_{n_0+n}$) and $f_{n,Z}^{\alpha}$ (with $Z_1, \ldots, Z_{n_0+n}$). Then*

$$\hat{\alpha} \text{ minimizes } -\frac{1}{n}\sum_{i=1}^{n} \log f_{n,X}^{\alpha,i}(X_i^\star) \;\Leftrightarrow\; \hat{\alpha} \text{ minimizes } -\frac{1}{n}\sum_{i=1}^{n} \log f_{n,Z}^{\alpha,i}(Z_i^\star).$$

Proof .
We will denote by $\{A_{i_1,\ldots,i_d}\}$ (resp. $\{B_{i_1,\ldots,i_d}\}$) the partition used to construct the modified histogram from the sample $X_1, \ldots, X_{n_0+n}$ (resp. $Z_1, \ldots, Z_{n_0+n}$). We have

$$\frac{1}{n}\sum_{i=1}^{n} \log f_{n,Z}^{\alpha,i}(Z_i^\star) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{n\mu_n^i\big(B(Z_i^\star)\big) + 1}{nh+1} g_\alpha(Z_i^\star)$$

where

$$g_\alpha(Z_i^\star) = \frac{1}{|det(Z(\alpha))|}\tilde{g}\big(Z(\alpha)^{-1}Z_i^\star + b_Z^\alpha\big)\,.$$

Theorem 2.2 and its proof give

$$\begin{cases} Z(\alpha) &= MX(\alpha) \\ b_Z^\alpha &= b_X^\alpha - \big(MX(\alpha)\big)^{-1}c \\ B_{i_1,\ldots,i_d} &= MA_{i_1,\ldots,i_d} + c\,. \end{cases}$$

Moreover, it is easily seen that $\mu_n^i\big(B(Z_i^\star)\big) = \mu_n^i\big(A(X_i^\star)\big)$. Putting all pieces together, we obtain

$$-\frac{1}{n}\sum_{i=1}^{n} \log f_{n,Z}^{\alpha,i}(Z_i^\star) = -\frac{1}{n}\sum_{i=1}^{n} \log f_{n,X}^{\alpha,i}(X_i^\star) + \log(|\det(M)|)\,.$$

Since M does not depend on α, the proof is complete.

■

4 Simulations

In this paragraph we are presenting some finite sample simulation results on the efficiency of the data-driven modified histogram f_n^α defined by (2.8) compared with the regular modified histogram f_n defined by (2.4). We use two data sets.
We first simulated 50 samples of size $n_0 + n$ ($n_0 + n = 150, 300, 550$) from bivariate normal populations with zero means, unit standard deviations and varying correlation coefficients $\rho = 0, 0.25, 0.5, 0.75$ and 0.95. For each sample, we have computed modified histograms f_n and f_n^α (α is selected by cross-validation). These estimates are built with

$$\begin{cases} \tilde{g}(x,y) = \exp\big(-x-\exp(-x)\big)\exp\big(-y-\exp(-y)\big) \\ n_0 = 50 \\ m_1 = m_2 = 4 \text{ for } n = 100 \\ m_1 = m_2 = 5 \text{ for } n = 250 \\ m_1 = m_2 = 6 \text{ for } n = 500\,. \end{cases}$$

We display in Table 2 the average of $D(f, f_n)$ and $D(f, f_n^\alpha)$ and the gain Ga in information divergence

$$Ga = \frac{D(f,f_n) - D(f,f_n^\alpha)}{D(f,f_n)}.$$

For the second set of data, points are generated from multivariate symmetric Laplace distributions (see Anderson, 1992) with density

$$f(x) = \frac{2}{(2\pi)^{d/2}|\Sigma|^{1/2}}(x^t\Sigma^{-1}x/2)^{v/2}K_v\Big(\sqrt{2x^t\Sigma^{-1}x}\Big)\,,$$

where $v = (2-d)/2$, Σ is a $d \times d$ non-negative definite symmetric matrix and $K_v(u)$ is the modified Bessel function of the third kind given by

$$K_v(u) = \frac{1}{2}\Big(\frac{u}{2}\Big)^v \int_0^\infty t^{-v-1}\exp\Big(-t-\frac{u^2}{4t}\Big)\,dt\,,\ u>0\,.$$

We set $d = 2, 4, 8, 10$ and several sample sizes $n_0 + n$. For each $(d, n_0 + n)$,

Table 2: Comparison of performance between regular and data-driven modified histograms.

n	ρ	$D(f, f_n)$	$D(f, f_n^\alpha)$	Ga
100	0	0.36	0.23	0.36
250		0.32	0.15	0.53
500		0.29	0.11	0.62
100	0.25	0.36	0.24	0.33
250		0.32	0.15	0.53
500		0.30	0.11	0.63
100	0.5	0.39	0.23	0.41
250		0.35	0.15	0.57
500		0.31	0.11	0.65
100	0.75	0.50	0.23	0.54
250		0.41	0.15	0.63
500		0.36	0.10	0.72
100	0.95	0.85	0.24	0.72
250		0.73	0.14	0.81
500		0.64	0.11	0.83

we simulated 50 samples from a symmetric Laplace distribution with

$$\Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \dots & \rho & 1 \end{pmatrix} \quad \rho = 0; 0.5; 0.95\,.$$

The density $\tilde{g}$ is a multivariate standard normal distribution and we take $m_j = 3$ $(j = 1, \dots, d)$. For $d = 2$ and 4 we again take $n_0 = 50$ to select α by cross-validation. However, for higher dimension the optimization problem is very heavy and takes too much time to reach an adequate solution. Thus, for $d = 8$ and 10, we propose the following alternative. We choose the transformation matrix $X(\alpha)$ in such a way that the image of $X_1, \dots, X_{n_0}$ has the same variance-covariance matrix as the density $\tilde{g}$ (identity in our example). In other words, we replace $X(\alpha)$ with $\hat{\Sigma}^{1/2}$ where $\hat{\Sigma}$ is an affine equivariant estimate of the variance-covariance matrix of the distribution (computed from $X_1, \dots, X_{n_0}$). The rest of the construction does not change. Note that the corresponding estimate no longer depends on α but on $\hat{\Sigma}$. In this regard it will be denoted by $f_n^{\hat{\Sigma}}$ and for the sake of clarity the associated reference density g_α and vector b_X^α will be denoted by $g_{\hat{\Sigma}}$ and $b_X^{\hat{\Sigma}}$. Consistency Corollary 2.1 is still true for $f_n^{\hat{\Sigma}}$. We take $n_0 = 1000$ for $d = 8$ and 10.

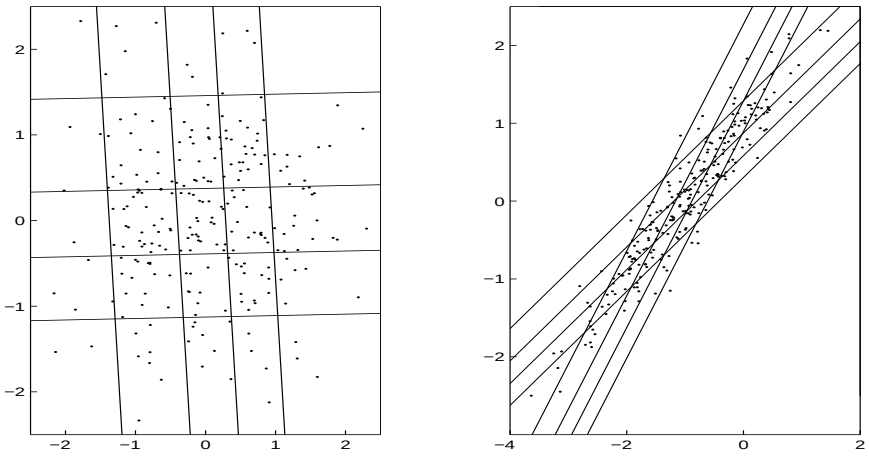

Figure 5: Partition of data-driven modified histogram. Simulated samples are the same as in Figure 3. $\tilde{g}$ is a product of Gumbel densities and $m_1 = m_2 = 5$.

$D(f, f_n)$ and $D(f, f_n^\alpha)$ are computed from Monte-Carlo method. The results are displayed in Table 3.

5 Concluding Remarks

1) Our examples demonstrate rather strikingly that f_n^α is on the whole better than f_n. The difference increases with the correlation and the dimension. Moreover, keeping n and d fixed, $D(f, f_n^\alpha)$ is stable whatever the correlation. It is worth pointing out that the partition is well adapted to the data cloud even with high correlation (see Figure 5).

2) Unlike with the first set of data, f_n^α is not better than f_n when $\rho = 0$ for the second set. It is due to the fact that the symmetric Laplace distribution and the standard gaussian distribution have the same median. Therefore the translation vector is close to zero and the reference density and the density to estimate are close enough without the transformation. On the other hand, for the first data set the two distributions do not have the same coordinatewise median. The translation can be seen as a "bias corrector" between the two densities.

3) For the second data set, the partition is not equivariant by affine transformation of the data sample when $d = 8$ or $d = 10$. All the same, the estimate is performant and the computation is quick even in large dimension. We emphasize that the transformation-retransformation procedure just allows to select a reference density which is not "too far" from the density to estimate.

Table 3: Comparison of performance between regular and data-driven modified histograms.

$d; n$	ρ	$D(f, f_n)$	$D(f, f_n^{\alpha})$ or $D(f, f_n^{\hat{\Sigma}})$	Ga
2;250	0	0.12	0.12	0
	0.5	0.18	0.13	0.28
	0.95	0.73	0.12	0.84
4;1000	0	0.34	0.36	-0.06
	0.5	0.55	0.37	0.33
	0.95	2.32	0.37	0.84
8;10000	0	0.90	0.90	0
	0.5	1.58	0.93	0.41
	0.95	6.10	0.93	0.85
10;500000	0	1.08	1.12	-0.04
	0.5	1.90	1.14	0.40
	0.95	7.52	1.12	0.85

In other words our choice of the transformation matrix is motivated by the fact that the reference density should be as close as possible to the density f to estimate . Since f is unknown in practice, we select the affine transformation such that the image of $X_1, \ldots, X_{n_0}$ and the random variable with density $\tilde{g}$ have the same variance-covariance matrix (with the help of the linear transformation $\hat{\Sigma}^{-1/2}$) and the same median (by the vector translation $b_X^{\hat{\Sigma}}$). Note that when f is elliptically symmetric, similar conditions on the choice of the transformation matrix are discussed by Chakraborty (2001) in the asymptotic study of the affine equivariant multivariate quantiles.

Acknowledgments. The authors wish to express their thanks to the referee for his comments and suggestions.

References

[1] D. N Anderson. A multivariate linnik distribution. *Statistic and Probability Letters*, 14:333–336, 1992.

[2] A. R Barron. The convergence in information of probability density estimators. In *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe : Japan, June 19-24 1988.

[3] A. R Barron, L Györfi, and E. C van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transaction on Information Theory*, 38:1437–1454, 1992.

[4] A. Berlinet and G. Biau. Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, 16:385–401, 2004.

[5] A. Berlinet and E. Brunel. Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, 16:493–513, 2004.

[6] A Berlinet, L Györfi, and E. C van der Meulen. The asymptotic normality of relative entropy in multivariate density estimation. *Publication de l'Institut de Statistique de l'Université de Paris*, 41:3–27, 1997.

[7] A Berlinet, I Vajda, and E. C van der Meulen. About the asymptotic accuracy of barron density estimates. *IEEE Transactions on Information Theory*, 38:1437–1454, 1998.

[8] B. M Brown. Statistical use of the spatial median. *Journal of the Royal Statistical Society, Series B*, 45:25–30, 1983.

[9] B .M Brown and T.P Hettmansperger. Affine invariant rank methods in the bivariate location model. *Journal of the Royal Statistical Society, Series B*, 49:301–310, 1987.

[10] B Chakraborty. On affine equivariant multivariate quantiles. *The Institute of Statistical Mathematics*, 53:380–403, 2001.

[11] P Chaudhuri. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91:862–872, 1996.

[12] P Chaudhuri and D Sengupta. Sign tests in multidimension : Inference based on the geometry of the data cloud. *Journal of the American Statistical Association*, 88:1363–1370, 1993.

[13] I Csizár. Information-type measures of divergence of probability distributions and indirect observations. *Studia Sci. Math. Hungar*, 2:299–318, 1967.

[14] I Csizár. Generalized entropy and quantization problems. In *Trans. Sixth Prague Conf. Information Theory, Statistical Decision Functions, Random Process*, pages 159–174, Prague : Academia, 1973.

[15] L. Devroye. On arbitrary slow rates of global convergence in density estimation. *Wahscheinlichkeitstheorie und verwandte Gebiete*, 62:475–483, 1983.

[16] W .F Eddy. Set valued ordering of bivariate data. In R.V. Ambartsumian and W. Weil, editors, *Stochastic Geometry, Geometric Statistics, and Stereology*, pages 79–90. Leipzig, 1983.

[17] W .F Eddy. Ordering of multivariate data. In L Billard, editor, *Computer Science and Statistics : The Interface*, pages 25–30. Amsterdam : North-Holland, 1985.

[18] L Györfi, F Liese, I Vajda, and E. C van der Meulen. Distribution estimates consistent in χ^2-divergence. *Statistics*, 32:31–57, 1998.

[19] P Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15:1491–1519, 1987.

[20] J. H. B Kemperman. On the optimum rate of transmitting information. *The Annals of Mathematical Statistics*, 40:2156–2177, 1969.

[21] J. H. B Kemperman. The median of a finite measure on a Banach space. In *Statistical Data Analysis Based on L_1 norm and Related Methods*, pages 217–230. Y. Dodge, Amsterdam North-Holland, 1987.

[22] S Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13:126–127, 1967.

[23] R. Y Liu, J. M Parelius, and K Singh. Multivariate analysis by data depth : descriptive statistics, graphics and inference (with discussion). *The Annals of Statistics*, 18:783–858, 1999.

[24] C. J Stone. An asymptotically optimal histogram selection rule. In L. Le Cam and R. A. Olshen, editors, *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 513–520, Wadsworth, Belmont, CA, 1985.

[25] J. W Tuckey. Mathematics and picturing data. In *Proc. Intern. Congr. Math*, volume 2, pages 523–531, Vancouver 1974, 1975.