ESAIM: PS
February 2006, Vol. 10, p. 24–45
DOI: 10.1051/ps:2006001

ESAIM: Probability and Statistics

HOW MANY BINS SHOULD BE PUT IN A REGULAR HISTOGRAM

Lucien Birgé[1] and Yves Rozenholc[2]

Abstract. Given an n-sample from some unknown density f on $[0,1]$, it is easy to construct an histogram of the data based on some given partition of $[0,1]$, but not so much is known about an optimal choice of the partition, especially when the data set is not large, even if one restricts to partitions into intervals of equal length. Existing methods are either rules of thumbs or based on asymptotic considerations and often involve some smoothness properties of f. Our purpose in this paper is to give an automatic, easy to program and efficient method to choose the number of bins of the partition from the data. It is based on bounds on the risk of penalized maximum likelihood estimators due to Castellan and heavy simulations which allowed us to optimize the form of the penalty function. These simulations show that the method works quite well for sample sizes as small as 25.

Mathematics Subject Classification. 62E25, 62G05.

Received July 7, 2003. Revised September 1, 2004 and May 11, 2005.

1. Introduction

Among the numerous problems that have been considered for a long time in Statistics, a quite simple one is: "How many bins should be used to build a *regular histogram*?" Here, by regular histogram, we mean one which is based on a partition into intervals of equal length. One can of course argue that this question is of little relevance nowadays since the histogram is an old fashioned estimator and that much more sophisticated and better methods are now available, such as variable bandwidths kernels or all kinds of wavelets thresholding. This is definitely true. Nevertheless, histograms are still in wide use and one can hardly see any other density estimator in newspapers. It is by far the simplest density estimator and probably the only one that can be taught to most students that take Statistics at some elementary level. And when you teach regular histograms to students, you unavoidably end up with the same question: "how many bins?". The question was also repeatedly asked to one of the authors by colleagues who were not professional mathematicians but still did use histograms or taught them to students.

Unfortunately, when faced to such a question, the professional statistician has no definitive answer or rather he has too many ones in view of the number of methods which have been suggested in the literature, none of them being completely convincing in the sense that it has been shown to be better than the others. The purpose of this paper is to propose a new fully automatic and easy to implement method to choose the number of bins to be used for building a regular histogram from data. The procedure, which we derived from a mixture

Keywords and phrases. Regular histograms, density estimation, penalized maximum likelihood, model selection.

[1] UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris VI, 4 Place Jussieu, 75252 Paris Cedex 05, France; lb@ccr.jussieu.fr;

[2] MAP5-UMR CNRS 8145, Université Paris 5, 45 rue des Saints-Pères, 75270 Paris Cedex 06, France; yves.rozenholc@math-info.univ.paris5.fr

of theoretical and empirical arguments, is not based on any smoothness assumptions and works quite well for all kinds of densities, even discontinuous, and sample sizes as small as 25. One could reasonably argue that our binwidth selection method, which requires the computation of many histograms, is substantially more complicated than the histogram itself and that its theoretical justifications are hard to explain to beginners. This is definitely true, but, as explained by Wand [26] the main point is the simplicity of the estimator itself so that any statistician or practitioner can understand the result. The binwidth selection procedure is typically viewed as a black box hidden in the statistical package that should be fast and statistically efficient. As such, our method has the advantage to be easy to implement and not time consuming.

One could also wonder why we restrict this study to regular histograms and we do not consider those based on irregular partitions, which clearly have a smaller bias for the same number of bins. There are two reasons for that. The first one is the search for practical efficiency. Optimizing the number of bins among regular partitions is computationally easy and fast. Considering all possible partitions involves a much more difficult optimization problem which requires much heavier and more delicate computations and sophisticated search procedures like dynamic programming. The second reason is of a more theoretical and fundamental nature. While the use of irregular partitions reduces the bias, the selection procedure involves a much larger number of choices which dramatically increases the complexity of the selection problem. To be more specific, there is only one regular partition of $[0,1]$ with D pieces, but the number of such partitions with endpoints on the grid $\{j/N, 0 \leq j \leq N\}$ is $\begin{pmatrix} N-1 \\ D-1 \end{pmatrix}$. This increase in the complexity results in a parallel increase of the random component of the risk, which is typically multiplied by a factor $\log N$, as shown by Proposition 2 of Birgé and Massart [4]. Therefore, the benefit due to bias reduction is often destroyed by this increase of the other component of the risk. We shall provide, at the end of this paper, a simulation study supporting these theoretical arguments. For this reason one should definitely not use irregular partitions systematically and deciding when one should introduce them or not is an even more delicate problem. Restricting to regular partitions appears to be a good compromise between speed and statistical efficiency.

There have been many attempts in the past to solve the problem of choosing an optimal number of bins from the data and we shall recall some of them in Section 4.1 below. Let us just mention here that, apart from some rules of thumbs like Sturges' rule (take approximately $1 + \log_2 n$ bins) or recommendations of the type: "one should have at least k observations in each cell" (k depending on the author), the methods we know are based on some asymptotic considerations. Rules of thumbs are very simple and do not aim at any optimality property. More sophisticated rules are based on the minimization of some asymptotic estimate of the risk. This is the case of methods like cross-validation or those based on the evaluation of the asymptotically optimal binwidth under smoothness assumptions for the underlying density. Methods connected with penalized maximum likelihood estimation, like Akaike's criterion or rules based on stochastic complexity or minimum description length are also derived from asymptotic considerations. It follows that the main drawback of all these rules is their asymptotic nature which does not warrant good performance for small sample sizes. Moreover, many of them are based on prior smoothness assumptions about the underlying density.

Our estimator is merely a generalization of Akaike's. This choice was motivated by some considerations about the *nonasymptotic* performances of penalized maximum likelihood estimators derived by Barron, Birgé and Massart [2]. For the specific case of histogram estimators, their results have been substantially improved by Castellan [5,6] and our study is based on her theoretical work. Roughly speaking, she has shown that a suitably penalized maximum likelihood estimator provides a data-driven method for selecting the number of bins which results in an optimal value of the Hellinger risk, up to some universal constant κ. The proof does not require any smoothness assumption and allows to consider discontinuous densities. Unfortunately, although Castellan's study indicates which penalty structure is suitable to get such a risk bound, theoretical studies are not powerful enough to derive a precise penalty function that would minimize the value of κ for small or moderate sample sizes.

In order to solve this problem, we performed an extensive simulation study including a large variety of densities and sample sizes in order to determine by an optimization procedure a precise form of the penalty function leading to a small value of κ. The resulting estimator is as follows. Assume that we have at disposal an n-sample $X_1, \ldots, X_n$ from some unknown density f (with respect to Lebesgue measure) on $[0, 1]$ and we want to design an histogram estimator $\hat{f}_D$ based on some partition $\{I_1, \ldots, I_D\}$ of $[0, 1]$ into D intervals of equal length. We choose for D the value $\hat{D}(X_1, \ldots, X_n)$ which maximizes $L_n(D) - \mathrm{pen}(D)$ for $1 \leq D \leq n/\log n$, where

$$L_n(D) = \sum_{j=1}^{D} N_j \log(DN_j/n) \quad \text{with } N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(X_i), \tag{1.1}$$

is the log-likelihood of the histogram with D bins and the penalty $\mathrm{pen}(D)$ is given by

$$\mathrm{pen}(D) = D - 1 + (\log D)^{2.5} \quad \text{for } D \geq 1. \tag{1.2}$$

This selection rule is simple enough to be explained to undergraduate students and easy to program, even by beginners, since one merely has to compute, for each value of D, the penalized likelihood. The resulting estimator $\hat{f}_{\hat{D}}$ will, from now on, be denoted $\tilde{f}_1$ and a few simulation results describing its performances graphically are given in Figure 1.

The previous construction is suitable for densities with a known compact support, which we may always assume to be $[0, 1]$ after a proper affine transform of the data. Often, the support of the density to estimate is unknown and even possibly noncompact. We therefore have to estimate a "pseudo-support". We adopt here the simplest strategy, taking for our estimated support the range of the data and performing a proper rescaling to change this range to $[0, 1]$.

In order to test our procedure, we first conducted another simulation study involving a large family of densities, sample sizes ranging from 25 to 1000 and different loss functions, to compare our method with a number of existing ones and the "oracle" which serves as a benchmark. Then, following the suggestions of a referee, we extended this comparison study to a few classical densities like the normal, exponential and uniform, extending the sample size up to 10 000.

The conclusion of this large scale empirical study, which is given in Section 4, is that among all of them, five estimators, $\tilde{f}_j$, with $1 \leq j \leq 5$ clearly outperform the others. Apart from our own estimator, these are based on $\mathbb{L}_2$ and Kullback crossvalidation, Akaike AIC criterion and the minimization of a stochastic complexity criterion. The final conclusion is that, on the average, our method essentially outperforms the others, although one of them, namely the one based on Rissanen's minimum complexity ideas (Rissanen [20]) and introduced, in the context of histogram estimation, by Hall and Hannan [13], is almost as good, in many cases. This is not so surprising since Rissanen's method and our approach are based on similar theoretic arguments.

The next section recalls the theoretical grounds on which our method is based while Section 3 describes the details of our simulation study. The results of the comparison with previous methods are given in Section 4. The Appendix contains some additional technical details.

2. Some theoretical grounds

2.1. Histograms and oracles

Let us first describe more precisely what is the mathematical problem to be solved. Let $X_1, \ldots, X_n$ be an n-sample from some unknown distribution with density f with respect to Lebesgue measure on some known compact interval which we assume, without loss of generality, to be $[0, 1]$. The histogram estimator of f based on the *regular partition with D pieces*, *i.e.* the partition $\mathcal{I}_D$ of $[0, 1]$ consisting of D intervals $I_1, \ldots, I_D$ of equal length $1/D$ is given by

$$\hat{f}_D = \hat{f}_D(X_1, \ldots, X_n) = \frac{D}{n} \sum_{j=1}^{D} N_j \mathbb{1}_{I_j}, \quad \text{with } N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(X_i). \tag{2.1}$$

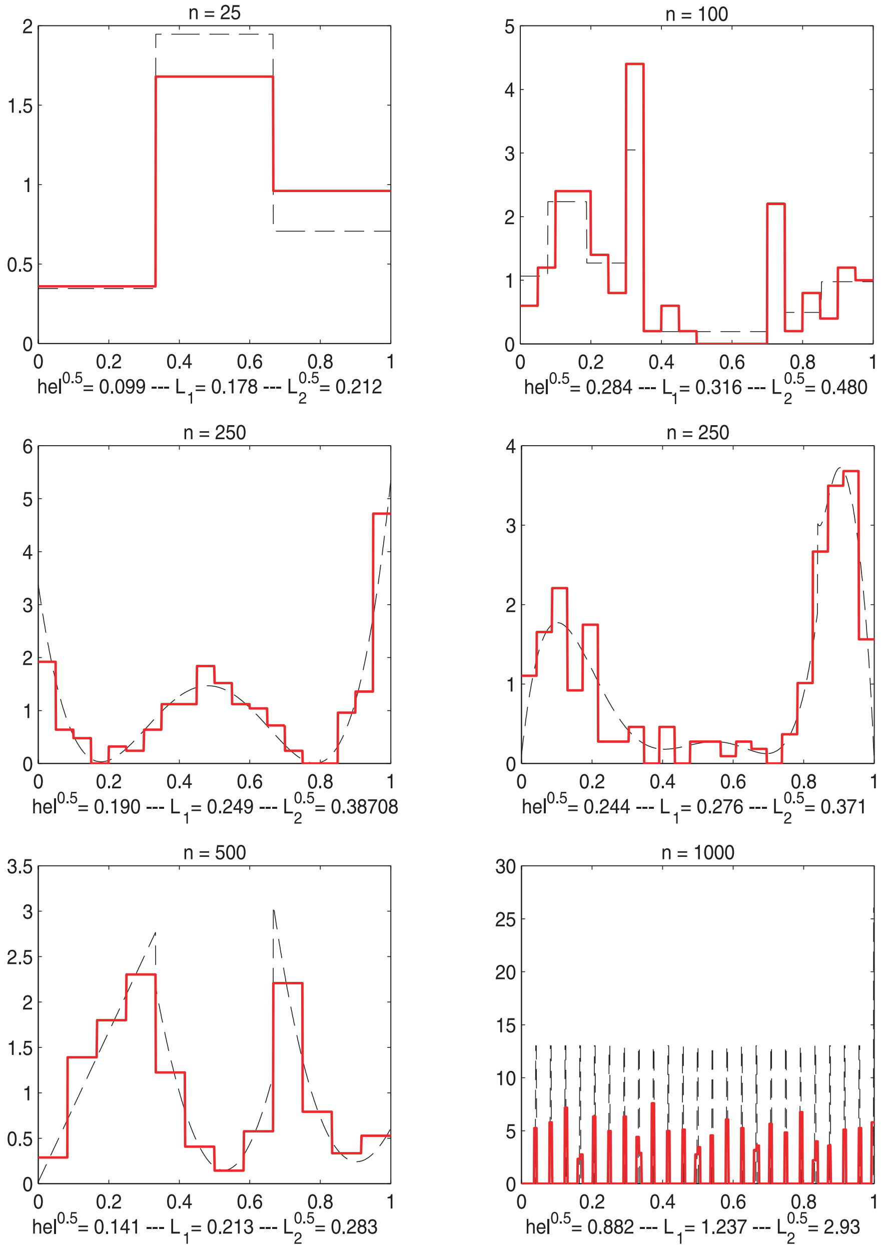

FIGURE 1. 6 examples, the thin mixed line represents the true density while the thick continuous line represents the estimator.

It is probably the oldest and simplest nonparametric density estimator. It is called the *regular histogram with D pieces* and it is the maximum likelihood estimator with respect to the set of piecewise constant densities on $\mathcal{I}_D$. In order to measure the quality of such an estimator, we choose some loss function ℓ and compute its risk

$$R_n(f,\hat{f}_D,\ell) = \mathbb{E}_f\left[\ell\left(f,\hat{f}_D(X_1,\ldots,X_n)\right)\right]. \tag{2.2}$$

From this decision theory point of view, the optimal value $D^{opt} = D^{opt}(f,n)$ of D is given by $R_n(f, \hat{f}_{D^{opt}}, \ell) = \inf_{D\geq 1} R_n(f, \hat{f}_D, \ell)$.

Unfortunately, no genuine statistical procedure can tell us what is the exact value of $D^{opt}(f,n)$ because it depends on the unknown density f to be estimated. This is why the procedure $\hat{f}_{D^{opt}}$ is called an *oracle* and $R_n(f, \hat{f}_{D^{opt}}, \ell)$ is the *risk of the oracle*. Obviously, an oracle is of no practical use but its risk can serve as a benchmark to evaluate the performance of any genuine data driven selection procedure $\hat{D}(X_1, \ldots, X_n)$. If $\hat{f}_{\hat{D}}$ denotes the histogram estimator based on the regular partition with $\hat{D}$ pieces, the quality of such a procedure at f can be measured by the value of the ratio

$$\frac{R_n(f, \hat{f}_{\hat{D}}, \ell)}{R_n(f, \hat{f}_{D^{opt}}, \ell)} = \frac{R_n(f, \hat{f}_{\hat{D}}, \ell)}{\inf_{D\geq 1} R_n(f, \hat{f}_D, \ell)}, \tag{2.3}$$

provided that the denominator is positive which requires to exclude the case of f being the uniform density, since, when $f = \mathbb{1}_{[0,1]}$, $D^{opt} = 1$, $\hat{f}_1 = f$ and $R_n(f, \hat{f}_1, \ell) = 0$.

Ideally, one would like this ratio to be bounded, uniformly with respect to f, by some constant C_n tending to one when n goes to infinity, and that $C_n - 1$ stay reasonably small even for moderate values of n. This is unfortunately impossible, as we just mentioned, since (2.3) is infinite when $f = \mathbb{1}_{[0,1]}$. By a continuity argument, $R_n(f, \hat{f}_1, \ell)$ may still be very small if f is very close to $\mathbb{1}_{[0,1]}$, which requires to exclude densities which are too close to the uniform – see the precise condition (3.1) below – from a comparison study based on the oracle criterion (2.3). A more precise discussion of this problem in the context of Gaussian frameworks can be found in Section 2.3.3 of Birgé and Massart [4].

2.2. **Loss functions**

Clearly, the value of D^{opt} and the performances of a given selection procedure $\hat{D}$ depend on the choice of the loss function ℓ. Popular loss functions include powers of $\mathbb{L}_p$-norms, for $1 \leq p < +\infty$ or the $\mathbb{L}_\infty$-norm, *i.e.*

$$\ell(f,g) = \|f-g\|_p^p \qquad \text{or} \qquad \ell(f,g) = \|f-g\|_\infty,$$

(with a special attention given to the cases $p = 1$ and 2), the squared Hellinger distance

$$h^2(f,g) = \frac{1}{2}\int_0^1 \left(\sqrt{f(y)} - \sqrt{g(y)}\right)^2 \mathrm{d}y,$$

and the Kullback-Leibler divergence (which is not a distance and possibly infinite) given by

$$K(f,g) = \int_0^1 \log\left(\frac{f(y)}{g(y)}\right) f(y)\,\mathrm{d}y \leq +\infty.$$

This last loss function is definitely not suitable to judge the quality of classical histograms since, as soon as $D \geq 2$, there is a positive probability that one of the intervals I_j be empty, implying that $K(f, \hat{f}_D) = +\infty$. A similar problem occurs with $K(\hat{f}_D, f)$ when f is not bounded away from 0.

Since we want to be able to deal with discontinuous densities f, the $\mathbb{L}_\infty$-norm is also inappropriate as a loss function since discontinuous functions cannot be properly approximated by piecewise constant functions on fixed partitions in $\mathbb{L}_\infty$-norm. By continuity, large values of p should also be avoided and we shall restrict ourselves hereafter to Hellinger distance and $\mathbb{L}_p$-norms for moderate values of p.

The most popular loss function in our context is probably the squared $\mathbb{L}_2$-loss for the reason that it is more tractable. Indeed,

$$\mathbb{E}_f\left[\left\|f - \hat{f}_D\right\|^2\right] = \mathbb{E}_f\left[\left\|\bar{f}_D - \hat{f}_D\right\|^2\right] + \left\|f - \bar{f}_D\right\|^2, \tag{2.4}$$

where $\bar{f}_D$ denotes the orthogonal projection (in the $\mathbb{L}_2$ sense) of f onto the D-dimensional linear space generated by the functions $\{\mathbb{1}_{I_j}\}_{1\leq j\leq D}$. In this case, the risk is split into a stochastic term and a bias term which may be analyzed separately. This accounts for the fact that optimizing the squared $\mathbb{L}_2$-risk of histogram estimators has been a concern of many authors, in particular Scott [22], Freedman and Diaconis [11], Daly [7], Wand [26] and Birgé and Massart [3].

Since the distribution of any selection procedure $\hat{D}(X_1,\ldots,X_n)$ only depends on the underlying distribution of the observations, it seems natural to evaluate its performances by a loss function which does not depend on the choice of the dominating measure. This is why some authors favour the systematic use of $\mathbb{L}_1$-loss for its nice invariance properties as explained by Devroye and Györfi [9], p. 2, and the preface of Devroye [8].

Although it is less popular, maybe because of its more complicated expression, we shall use here the squared Hellinger distance as our reference loss function to determine a suitable penalty, this choice being actually based on theoretical grounds only. First, as the $\mathbb{L}_1$-distance, it is a distance between probabilities, not only between densities. Then it is known that it is the natural distance to use in connection with maximum likelihood estimation and related procedures, as demonstrated many years ago by Le Cam (see, for instance, Le Cam [18] or Le Cam and Yang, [19]). Finally, the results of Castellan [5,6] that we use here are based on it.

Of course, the choice of a “nice” loss function is, for a large part, a question of personal taste. Hellinger distance has already been used as loss function in the context of regular histogram density estimation by Kanazawa [17] but other authors do prefer $\mathbb{L}_p$-losses and one should read for instance the arguments of Devroye mentioned above or those of Jones [16]. Therefore, although we shall base our choice of the procedure $\hat{D}$ on the Hellinger loss, we shall also use other loss functions, including squared $\mathbb{L}_2$, to evaluate its performances and compare it to other methods.

2.3. Hellinger risk

Before we consider the problem of choosing an optimal value of D we need an evaluation of the risk of regular histograms $\hat{f}_D$ for a given value of D to compute the ratio (2.3). There is actually nothing special with regular histograms from this point of view and we shall consider arbitrary partitions in this section. Given an histogram estimator $\hat{f}_{\mathcal{I}}$ of the form (2.5) below based on some arbitrary partition $\mathcal{I}$, its risk is given by $\mathbb{E}_f\left[h^2(f,\hat{f}_{\mathcal{I}})\right]$. Asymptotic evaluations of this risk are given by Castellan [5,6] and a nonasymptotic bound is as follows. Since we were unable to find this result in the literature, we include the proof, which follows classical lines, in the Appendix, for completeness.

Theorem 1. *Let f be some density with respect to some measure μ on $\mathcal{X}$, $X_1,\ldots,X_n$ be an n-sample from the corresponding distribution and $\hat{f}_{\mathcal{I}}$ be the histogram estimator based on some partition $\mathcal{I}=\{I_1,\ldots,I_D\}$ of $\mathcal{X}$, i.e.*

$$\hat{f}_{\mathcal{I}}(X_1,\ldots,X_n)=\frac{1}{n}\sum_{j=1}^{D}\frac{N_j}{\mu(I_j)}\mathbb{1}_{I_j},\quad \textit{with } N_j=\sum_{i=1}^{n}\mathbb{1}_{I_j}(X_i). \tag{2.5}$$

Setting $p_j=\int_{I_j}f\mathrm{d}\mu$, we get

$$\mathbb{E}_f\left[h^2(f,\hat{f}_{\mathcal{I}})\right]\leq h^2(f,\bar{f}_{\mathcal{I}})+\frac{D-1}{2n}\quad \textit{with } \bar{f}_{\mathcal{I}}=\sum_{j=1}^{D}\frac{p_j}{\mu(I_j)}\mathbb{1}_{I_j}. \tag{2.6}$$

Moreover

$$\mathbb{E}_f\left[h^2(f,\hat{f}_{\mathcal{I}})\right]=h^2(f,\bar{f}_{\mathcal{I}})+\frac{D-1}{8n}[1+o(1)], \tag{2.7}$$

when $n(\inf_{1\leq j\leq D}p_j)$ tends to infinity.

Remark. It should be noticed that $\bar{f}_{\mathcal{I}}$ minimizes both the $\mathbb{L}_2$-distance between f and the space $H_{\mathcal{I}}$ of piecewise constant functions on $\mathcal{I}$ and the Kullback-Leibler information number $K(f,g)$ between f and some element g of $H_{\mathcal{I}}$, but not the Hellinger distance $h(f,H_{\mathcal{I}})$. In the case of a regular partition, $\bar{f}_{\mathcal{I}}=\bar{f}_D$ as in (2.4).

2.4. Penalized maximum likelihood estimators

The theoretical properties of penalized maximum likelihood estimators over spaces of piecewise constant densities, on which our work is based, have been studied by Castellan [5,6]. We recall that a penalized maximum likelihood estimator derived from a penalty function $D \mapsto \mathrm{pen}(D)$ is the histogram estimator $\hat{f}_{\hat{D}}$ where $\hat{D}$ is a maximizer with respect to D of $L_n(D) - \mathrm{pen}(D)$ with $L_n(D)$ given by (1.1). Roughly speaking, Castellan's results say (not going into details in order to avoid technicalities) that one should use penalties of the form

$$\mathrm{pen}(D) = c_1(D-1)\left(1 + c_2\sqrt{L_D}\right)^2 \quad \text{with } c_1 > 1/2, \quad L_D > 0, \tag{2.8}$$

where c_2 is a suitable positive constant and the numbers L_D satisfy

$$\sum_{D \geq 1} \exp[-(D-1)L_D] = \Sigma < +\infty. \tag{2.9}$$

Let us observe that this family of penalties includes the classical Akaike's AIC criterion corresponding to $\mathrm{pen}(D) = D - 1$ (choose for instance $c_1 = 3/4$ and $L_D = L$ in a suitable way). Defining $\hat{D}(X_1, \ldots, X_n)$ as the maximizer of $L_n(D) - \mathrm{pen}(D)$ for $1 \leq D \leq \bar{D} = \Gamma n/(\log n)^2$, Castellan proves, under suitable assumptions (essentially that f is bounded away from 0), that

$$\mathbb{E}_f\left[h^2(f, \hat{f}_{\hat{D}})\right] \leq \kappa(c_1) \inf_{1 \leq D \leq \bar{D}} \left\{K(f, \bar{f}_D) + n^{-1}\mathrm{pen}(D)\right\} + n^{-1}(\Sigma + 1)\kappa', \tag{2.10}$$

where κ, κ' are positive constants, κ depending on c_1, κ' on the parameters involved in the assumptions. This bound and (2.9) suggest to choose some non-increasing sequence $(L_D)_{D\geq 1}$ leading to some Σ of moderate size, which we shall assume from now on.

The asymptotic evaluations of Castellan also suggest to choose $c_1 = 1$ in order to minimize $\kappa(c_1)$, at least when D^{opt} goes to infinity. In this case, the penalty given by (2.8) can be viewed as a modified AIC criterion with an additional correction term which warrants its good behaviour when the number of observations and therefore the number of cells to be considered in the partition, are not large. Both criteria are equivalent when D tends to infinity.

It is also known (see for instance Birgé and Massart [4], Lem. 5) that, when $\left|\log(f/\bar{f}_D)\right|$ is small, $K(f, \bar{f}_D)$ is approximately equal to $4h^2(f, \bar{f}_D)$. Therefore, under suitable assumptions on f, setting $c_1 = 1$, one can show, by the boundedness of the sequence $(L_D)_{D\geq 1}$, that

$$\mathbb{E}_f\left[h^2(f, \hat{f}_{\hat{D}})\right] \leq \kappa_1 \inf_{1 \leq D \leq \bar{D}} \left\{h^2(f, \bar{f}_D) + \frac{D-1}{n}\right\} + \kappa_2 \frac{1+\Sigma}{n}\cdot \tag{2.11}$$

In view of Theorem 1, one can finally derive from (2.11) that, under suitable restrictions on f and for n large enough,

$$\mathbb{E}_f\left[h^2(f, \hat{f}_{\hat{D}})\right] \leq \kappa_1' \mathbb{E}_f\left[h^2(f, \hat{f}_{D^{opt}})\right] + \kappa_2'/n. \tag{2.12}$$

3. From theory to practice

3.1. Some heuristics

Although the asymptotic considerations suggest to choose $c_1 = 1$ (and this was actually confirmed by our simulations), the theoretical approach is not powerful enough to indicate precisely how one should choose the sequence $(L_D)_{D\geq 1}$ in order to minimize the risk. It simply suggests that Σ should not be large in order to keep the remainder term $\kappa_2(\Sigma + 1)/n$ of moderate size when n is not very large. In order to derive a form of penalty that leads to a low value of the risk, one needs to perform an optimization based on simulations and, at this stage, some heuristics will be useful. In particular we shall pretend that the asymptotic formula (2.7) is exact, and use the approximation

$$\mathbb{E}_f\left[h^2(f,\hat{f}_D)\right] \approx h^2(f,\bar{f}_D) + (D-1)/(8n),$$

which implies that $\mathbb{E}_f\left[h^2(f,\hat{f}_D)\right] \gtrapprox (8n)^{-1}$ for $D \geq 2$. Together with (2.12), this implies that

$$\mathbb{E}_f\left[h^2(f,\hat{f}_{\hat{D}})\right] \leq \kappa_3 \mathbb{E}_f\left[h^2(f,\hat{f}_{D^{opt}})\right] \quad \text{for } D^{opt} > 1.$$

If $D^{opt} = 1$, this bound still holds provided that

$$8nh^2(f,\mathbb{1}_{[0,1]}) \geq 1, \tag{3.1}$$

since $\bar{f}_1 = \mathbb{1}_{[0,1]}$, whatever f, which means that f is not too close to the uniform. This restriction confirms the arguments of Section 2.1.

If $c_1 = 1$, (2.8) can be written

$$\text{pen}(D) = D - 1 + c_2(D-1)\left(2\sqrt{L_D} + c_2 L_D\right).$$

On the one hand, the constant c_2 is only known approximately and (2.9) requires that $(D-1)L_D$ tends to infinity with D. On the other hand, the risk bound (2.10) includes a term proportional to $n^{-1}\,\text{pen}(D)$, therefore L_D should not be large. A natural trade-off leads to consider penalties of the form $\text{pen}(D) = D - 1 + g(D)$ where the function g tends to infinity with D but not too fast. In order to keep the simulation time finite, we had to consider only some specific functions g and we actually restricted our search to three parametric families exhibiting various behaviours, namely:

$$\alpha x^{\beta},\ \ 0 < \beta < 1; \quad \alpha x(1+\log x)^{-\beta},\ \ \beta > 0 \quad \text{and} \quad \alpha(\log x)^{\beta},\ \ \beta > 1, \tag{3.2}$$

with $\alpha > 0$ in all three cases. We tried, through heavy simulations, to determine the best shape and some "optimal" values for α and β. These were found by successive discretizations of the parameters using finer grids at each step. We also replaced the restriction $1 \leq D \leq \bar{D}$ with $\bar{D} = \Gamma n/(\log n)^2$ with some constant $\Gamma > 0$ of Castellan's Theorem by the simpler condition $\bar{D} = n/\log n$ which did not lead to any trouble in practice.

3.2. The operational procedure

We proceeded in two steps. The first one was an *optimization step* to choose a convenient value for the function g, the second one was a *comparison step* to compare our new procedure with more classical ones. In both cases, we had to choose some specific densities to serve as references, *i.e.* for which we should evaluate the performances of the different estimators. For the optimization step, we chose densities of piecewise polynomial form. To define them we used either the partition with a single element which leads to continuous densities, or some regular or irregular partitions with several elements. Both the partitions and the coefficients of the polynomials given by their linear expansion within the Legendre basis were drawn using a random device which ensured positivity. We then removed from the initial family some elements which were too redundant and also added some special piecewise constant densities which were known to be difficult to estimate. We finally ended up with the set of 30 different densities which are shown in Figure 9 in the Appendix. Some of these densities are not smooth and this choice was made deliberately. Histograms are all purpose rough estimates which should cope with all kinds of densities. Testing their performances only with smooth densities like the normal or beta is not sufficient, from our point of view. When the densities were chosen, we selected a range of values for n, namely $n = 25, 50, 100, 250, 500$ and 1000 and this resulted in a set $\mathcal{F}$ of 176 pairs (f,n) after we excluded those for which the value of the oracle would be too close to zero, *i.e.* four pairs which did not satisfy the requirement (3.1) for the reasons explained in Section 2.1.

For the comparison step, we compared the performances of a selected set of other methods to the oracle, using again $\mathcal{F}$ as our test set. We then tested the other methods against ours on a few classical and possibly noncompactly supported densities, in this case estimating a pseudo-support as explained in Section 1.

In both steps of our simulation study (optimizing the penalty and comparing the resulting procedure with others), we had to evaluate risks $R_n(f, \tilde{f}, \ell)$ for various procedures $\tilde{f}$. There are typically no closed form formulas for such theoretical risks and we had to replace them by empirical risks based on simulations. We systematically used the same method: given the pair (f, n) we generated on the computer 1000 pseudo-random samples $X_1^j, \ldots, X_n^j$, $1 \leq j \leq 1000$ of size n and density f. We then performed all our computations replacing the theoretical distributions of losses of the procedures $\tilde{f}$ at hand: $\mathbb{P}_f[\ell(f, \tilde{f}(X_1, \ldots, X_n)) \leq t]$ by their empirical counterparts

$$\overline{\mathbb{P}}_n\left[\ell\left(f, \tilde{f}(X_1, \ldots, X_n)\right) \leq t\right] = \frac{1}{1000}\sum_{j=1}^{1000} \mathbb{1}_{[0,t]}\left(\ell\left(f, \tilde{f}(X_1^j, \ldots, X_n^j)\right)\right).$$

In particular we approximated the true risk $R_n(f, \tilde{f}, \ell)$ by its empirical version

$$\overline{R}_n(f, \tilde{f}, \ell) = \frac{1}{1000}\sum_{j=1}^{1000} \ell\left(f, \tilde{f}(X_1^j, \ldots, X_n^j)\right),$$

and the upper 95% quantile of the distribution of $\ell(f, \tilde{f}(X_1, \ldots, X_n))$ by the corresponding upper 95% quantile $\overline{Q}_{(0.95)}(n, f, \tilde{f}, \ell)$ of the empirical distribution $\overline{\mathbb{P}}_n$. Note here that such computations required the evaluations of quantities of the form $\ell(f, \tilde{f})$, namely $h^2(f, \tilde{f})$ or $\|f - \tilde{f}\|_p^p$. Since both f (piecewise polynomial) and $\tilde{f}$ (piecewise constant) were piecewise continuous, we could compute the losses by numerical integration separately on each of the intervals where both functions were continuous. The precise details of the procedures and the corresponding MATLAB functions can be found on the WEB site `http://www.proba.jussieu.fr/~rozen`.

3.3. The optimization

In this step, we wanted to compare the performances of the various penalized maximum likelihood estimators with penalties of the form $\mathrm{pen}(D) = D - 1 + g(D)$ according to the possible values of g over the testing class $\mathcal{F}$. The performance of a selection procedure $\hat{D}(g)$ based on the penalty involving some function g was evaluated by a comparison of its empirical risk with the empirical optimal risk corresponding to $D = D^{opt}$. For a given procedure $\tilde{f}$, a loss function ℓ and a testing pair (f, n) we denote the corresponding empirical risk ratio by

$$\overline{M}_n(f, \tilde{f}, \ell) = \frac{\overline{R}_n(f, \tilde{f}, \ell)}{\inf_{D \geq 1} \overline{R}_n(f, \hat{f}_D, \ell)}. \tag{3.3}$$

Ideally, one would like to minimize $\overline{M}_n(f, \hat{f}_{\hat{D}(g)}, h^2)$, with respect to g for all pairs $(f, n) \in \mathcal{F}$. Of course the optimal strategy depends on the pair and we looked for some uniform bound for $\overline{M}_n$ but it appeared that, roughly speaking, $\overline{M}_n$ behaves as a decreasing function of $D^{opt}(f, n)$ so that we tried to minimize approximately

$$\sup_{\{(f,n)\,|\,D^{opt}(f,n)=k\}} \overline{M}_n(f, \hat{f}_{\hat{D}(g)}, h^2),$$

with respect to g for all values of k simultaneously. Again, this is not a well-defined problem and, of course, no specific function g was found uniformly better than the others for all densities and all sample sizes so that some compromises were needed for the final choice. We just looked for a function g the performances of which were always reasonably close to the best ones for all densities in our test set and all values of n. This ruled out most parameter values (α, β) because most rules were found bad on some density or those which were good for small samples became bad for larger ones. We ended with only a few parameter values that were acceptable and a close inspection of the long lists of risk values determined our final choice $g(x) = (\log x)^{2.5}$, although some other close parameter values were found to lead to almost equivalent results.

3.4. The performances of our estimator

In order to evaluate the performances of $\tilde{f}_1$, we compared its risk with the oracle for all pairs $(f, n) \in \mathcal{F}$ and various loss functions. We actually considered, for all our comparisons, four typical loss functions, which are powers of either the Hellinger or some $\mathbb{L}_p$-distances. More precisely we set $p_0 = p_2 = 2$, $p_1 = 1$, $p_3 = 5$ and

$$\ell_0(f, f') = h^{p_0}(f, f') \quad \text{and} \quad \ell_i(f, f') = \|f - f'\|_{p_i}^{p_i} \quad \text{for } i = 1,\, 2,\, 3.$$

In order to facilitate comparisons between the different loss functions and to balance the effect of the differences in the powers p_i, we expressed our results in terms of a normalized version $\overline{M}_n^\star$ of $\overline{M}_n$, setting, for any density f and estimator $\tilde{f}$,

$$\overline{M}_n^\star(f, \tilde{f}, \ell_i) = \left[\overline{M}_n(f, \tilde{f}, \ell_i)\right]^{1/p_i}.$$

The results of the comparisons are summarized in Table 1 below. For each n, we denote by $\mathcal{F}_n$ the set of densities f such that $(f, n) \in \mathcal{F}$. For $n \geq 100$, $\mathcal{F}_n$ contains all 30 densities we started with, but four of them (three for $n = 25$ and one for $n = 50$), which are too close to the uniform, had to be excluded since they do not satisfy (3.1). Table 1 gives the values of $\sup_{f \in \mathcal{F}_n} \overline{M}_n^\star(f, \tilde{f}_1, \ell_i)$ for the different values of n and i. We see that all values of $\overline{M}_n^\star(f, \tilde{f}_1, h^2)$ are smaller than 1.5 and not much larger for the $\mathbb{L}_1$- and $\mathbb{L}_2$-losses, although our procedure was not optimized for those losses. Not surprisingly, the results for $\mathbb{L}_5$ are worse for small values of n but improve substantially for $n \geq 500$.

TABLE 1. Maximum normalized mean ratio: $\sup_{f \in \mathcal{F}_n} \overline{M}_n^\star(f, \tilde{f}_1, \ell_i)$.

$i \setminus n$	25	50	100	250	500	1000
0	1.40	1.38	1.43	1.30	1.30	1.26
1	1.48	1.54	1.49	1.34	1.33	1.26
2	1.84	1.64	1.49	1.48	1.42	1.38
3	2.94	2.89	2.85	2.55	1.62	1.53

We also give below in Figure 2 the values of $\log_2[\overline{M}_n^\star(f, \tilde{f}_1, \ell_i)]$ for all pairs $(f, n) \in \mathcal{F}$ and $0 \leq i \leq 3$. In each case we ordered our set $\mathcal{F}$ in increasing order of the empirical risk of the oracle $\inf_{D \geq 1} \overline{R}_n(f, \hat{f}_D, \ell_i)$ for the corresponding loss function. The binary logarithm of this risk is printed as a grey dashed line on the same graphic with its scale on the right side. This means that the right-hand side of the graph corresponds to those densities which cannot be estimated accurately, even by the oracle, with the given number of observations. It is interesting to notice that some values of $\overline{M}_n^\star(f, \tilde{f}_1, \ell_i)$, when n is large and f is a "nice" density, are actually smaller than one, which means that, under favorable circumstances, our estimator "beats" the oracle. This is an additional illustration of a known fact: the oracle has a fixed number D^{opt} of bins (the one which minimizes the risk, *i.e.* the average loss) independently of the sample, while a good selection procedure tries to optimize the number of bins for each sample and can therefore adjust to the peculiarities of the sample. This means that the very notion of an oracle is questionable as an absolute reference. It nevertheless remains a very convenient benchmark.

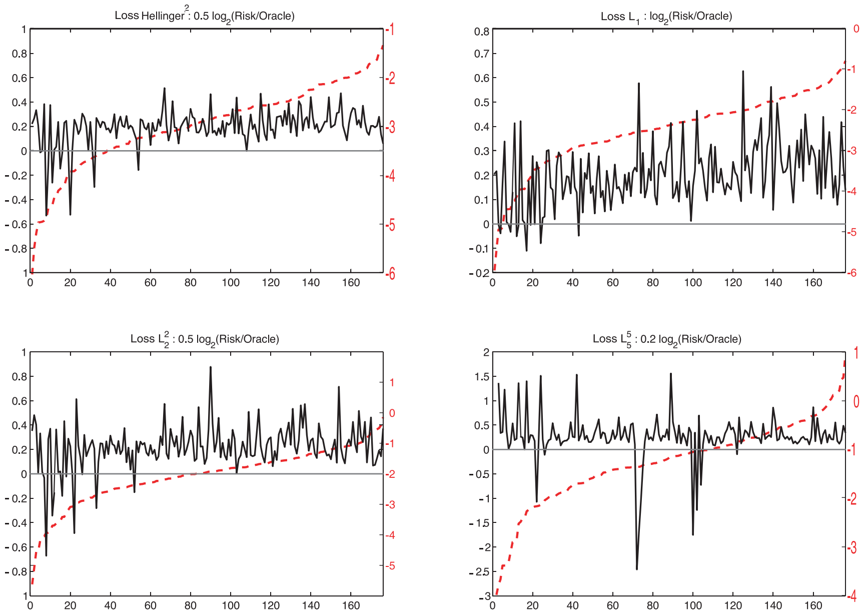

FIGURE 2. Normalized binary logarithm of the ratio between the risk of our method and the oracle (black line with scale reference on the left side) for all pairs $(f, n) \in \mathcal{F}$ sorted in increasing order of the normalized binary logarithm of the risk of the oracle (grey dashed line with scale reference on the right side) for the relevant losses.

4. Comparison with previous methods

4.1. Choosing some reference procedures

4.1.1. *Some historical remarks*

The following discussion is supposed to give only a brief summary of the set of methods in use to build histograms and does not pretend to be exhaustive in any way. We just selected a number of methods that, from our point of view, reflected the various approaches to binwidth selection and, as such, could make up a suitable set of reference procedures with which we could compare our method.

The first methods used to decide about the number of bins were just rules of thumbs and date back to Sturges [23]. According to Wand [26] such methods are still in use in many commercial softwares although they do not have any type of optimality property. Methods based on theoretical grounds appeared more recently and they can be roughly divided into three classes.

If the density to be estimated is smooth enough (has a continuous derivative, say), it is often possible, for a given loss function, to evaluate the optimal asymptotic value of the binwidth, the one which minimizes the risk, asymptotically. Such evaluations have been made by Scott [22] and Freedman and Diaconis [11] for the squared $\mathbb{L}_2$-loss, by Devroye and Györfi [9] for the $\mathbb{L}_1$-loss and by Kanazawa [17] for the squared Hellinger distance. Unfortunately, the optimal binwidth is asymptotically of the form $cn^{-1/3}$ where c is a functional of the unknown density to be estimated and its derivative. Many authors suggest to use the Normal rule to derive c. One can alternatively estimate it by a plug-in method as in Wand [26].

Methods based on cross-validation have the advantage to avoid the estimation of an asymptotic functional and directly provide a binwidth from the data. An application to histograms and kernel estimators is given in Rudemo [21]. Theoretical comparisons between Kullback cross-validation and the AIC criterion are to be found in Hall [12].

The third class of methods includes specific implementations for the case of regular histograms of general criteria used for choosing the number of parameters to put in a statistical model. The oldest method is the minimization of Akaike's AIC criterion (see Akaike [1]). Akaike's method is merely a penalized maximum likelihood method with penalty $\text{pen}(D) = D - 1$ in our case. In view of (1.2), our criterion is just a generalization of AIC criterion tuned for better performance with small samples. Taylor [24] derived the corresponding asymptotic optimal binwidth (under smoothness assumptions on the underlying density) which turns out to be the same as the asymptotically optimal binwidth for squared Hellinger risk, as derived by Kanazawa [17]. Related methods are those based on minimum description length and stochastic complexity due to Rissanen (see for instance Rissanen [20]). Their specific implementation for histograms has been discussed in Hall and Hannan [13].

Somewhat more exotic methods have been proposed by Daly [7] and He and Meeden [14], the second one being based on Bayesian bootstrap.

4.1.2. *The selected methods*

Following the previous historical review, we selected 13 different estimators $\tilde{f}_2, \ldots, \tilde{f}_{14}$ to compare with our own and evaluate the performances of our method. The precise description of these estimators and of their implementation can be found in the Appendix. Let us just briefly mention that $\tilde{f}_2$ and $\tilde{f}_3$ are respectively $\mathbb{L}_2$ and Kullback-Leibler cross-validation methods, $\tilde{f}_4$ is the minimization of AIC, $\tilde{f}_5$ and $\tilde{f}_6$ are based on stochastic complexity and minimum description length respectively, $\tilde{f}_7$ to $\tilde{f}_{10}$ are estimators based on asymptotic evaluations of an optimal binwidth according to various criteria, $\tilde{f}_{11}$ is Sturges'rule, $\tilde{f}_{12}$ is due to Daly and $\tilde{f}_{13}$ to He and Meeden. For completeness, we added $\tilde{f}_{14}$ which is due to Devroye and Lugosi and described in Section 10.3 of Devroye and Lugosi [10]. It does not look for the optimal regular partition but only for the optimal partition among dyadic ones.

We actually also studied the performances of modified versions of some of those estimates, as described by Rudemo [21], Hall and Hannan [13] and Wand [26]. Since the performances of the modified methods were found to be similar to or worse than those of the original estimators, we do not include them here.

4.2. **Our first comparison study**

We first computed, for all 176 pairs $(f, n) \in \mathcal{F}$, all estimators $\tilde{f}_k, 1 \leq k \leq 14$ and the four selected loss functions, the values of $\overline{M}^{\star}_n(f, \tilde{f}_k, \ell_i)$. This resulted in a large set of data which had to be summarized. Therefore, for each n, ℓ and k we considered the set $S(n, \ell, k)$ of the $|\mathcal{F}_n|$ values of $\overline{M}^{\star}_n(f, \tilde{f}_k, \ell)$ for $f \in \mathcal{F}_n$. A preliminary and rough comparison of the estimates can be based on the boxplots of the different sets $S(n, \ell, k)$. Here, the box provides the median and quartiles, the tails give the 10 and 90% quantiles and the 6 additional points (three on each side of the box) give the values which are outside this range.

Figure 3 shows the boxplots corresponding to all methods for $n = 25, 100$ and 1000, squared Hellinger and $\mathbb{L}_2^2$ losses. It is readily visible from these plots that estimators $\tilde{f}_6$ to $\tilde{f}_{14}$ are not satisfactory for $n \geq 100$ as compared to the others. We also used as a secondary index of performance the normalized ratio

$$\overline{Q}^{\star}_n(f, \tilde{f}, \ell_i) = \left[\frac{\overline{Q}_{(0.95)}(n, f, \tilde{f}, \ell_i))}{\inf_{D \geq 1} \overline{R}_n(f, \hat{f}_D, \ell_i)}\right]^{1/p_i}$$

of the empirical version of the 95% quantile of the distribution of $\tilde{f}$ to the corresponding risk of the oracle and drawn the corresponding boxplots. The results are quite similar to those of Figure 3. The complete set of results (not included here to avoid redundancy) shows that the performances of the estimators $\tilde{f}_6$ to $\tilde{f}_{14}$ for the other values of n and loss functions are not better.

For each of the four "good" estimators $\tilde{f}_j$, $2 \leq j \leq 5$ and each loss function ℓ_i, we provide hereafter in Figure 4 (one line for each loss function, one column for each $\tilde{f}_j$) the values of $\log_2[\overline{M}^{\star}_n(f, \tilde{f}_j, \ell_i)/\overline{M}^{\star}_n(f, \tilde{f}_1, \ell_i)]$ (*i.e.* the binary logarithms of the normalized ratio of the risks of the best four methods to the risk of our method

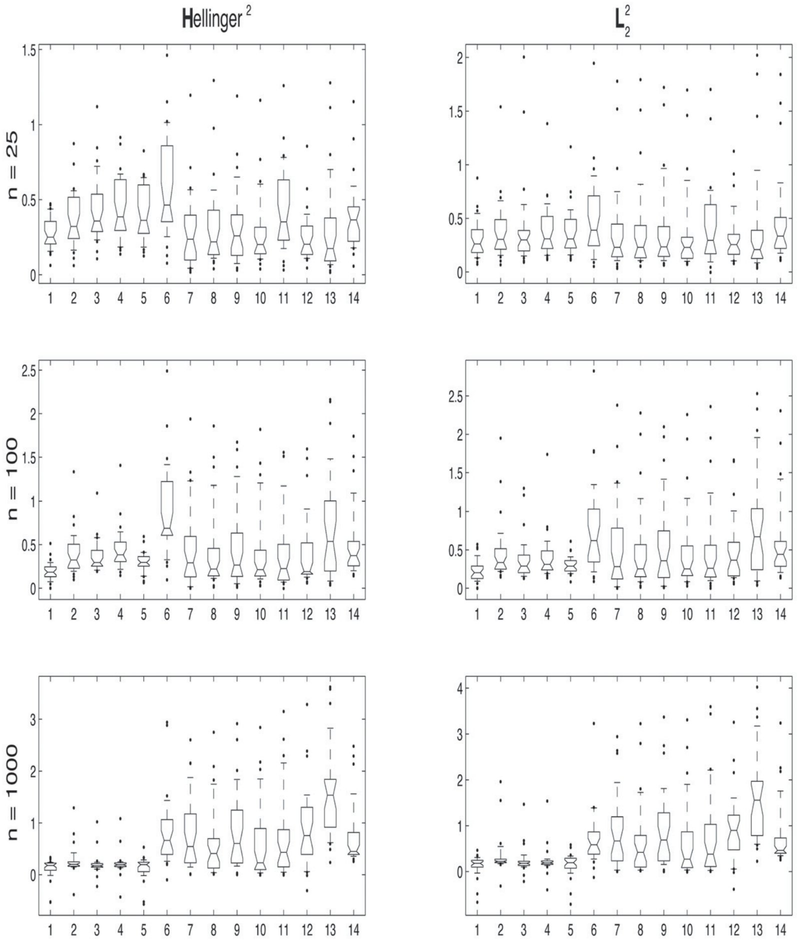

FIGURE 3. Boxplot of Hellinger^2 (left column) and $\mathbb{L}_2^2$ (right column) binary logarithm of the normalized ratio between the risk of the estimators and that of the oracle for $n = 25, 100, 1000$.

for the loss ℓ_i), for all pairs $(f, n) \in \mathcal{F}$. As for Figure 2, we ordered our set $\mathcal{F}$ in increasing order of the empirical risk of the oracle for the corresponding loss function and added this risk as a grey dashed line to serve as a reference. On these graphs, our method appears to be most of the time superior to the others, especially for the pairs (f, n) with a small value of the risk of the oracle, *i.e.* on the left side of the graphs. Some other methods do perform better on the extreme right of the graphs, for the pairs with numbers superior to 150. Such pairs correspond to densities that cannot be well estimated by any method with the number of available observations because the risk of the oracle is quite large for those densities.

We actually also drew the same graphics for the remaining estimators $\tilde{f}_j$ with $j \geq 6$ and the results confirmed what we already derived from the boxplots, namely that these estimators are clearly outperformed by the five first ones. To save space, we do not include these additional graphics.

As we previously noticed, some methods, corresponding to estimators $\tilde{f}_j$ with $6 \leq j \leq 14$ were found to behave rather poorly on our test set, especially for large n. This is actually not surprising for Sturges'rule ($\tilde{f}_{11}$) which is a rule of thumbs, for $\tilde{f}_{12}$ since the theoretical arguments supporting Daly's method are not very strong and for $\tilde{f}_{13}$ since the decision theory arguments used by He and Meeden involve a very special loss function

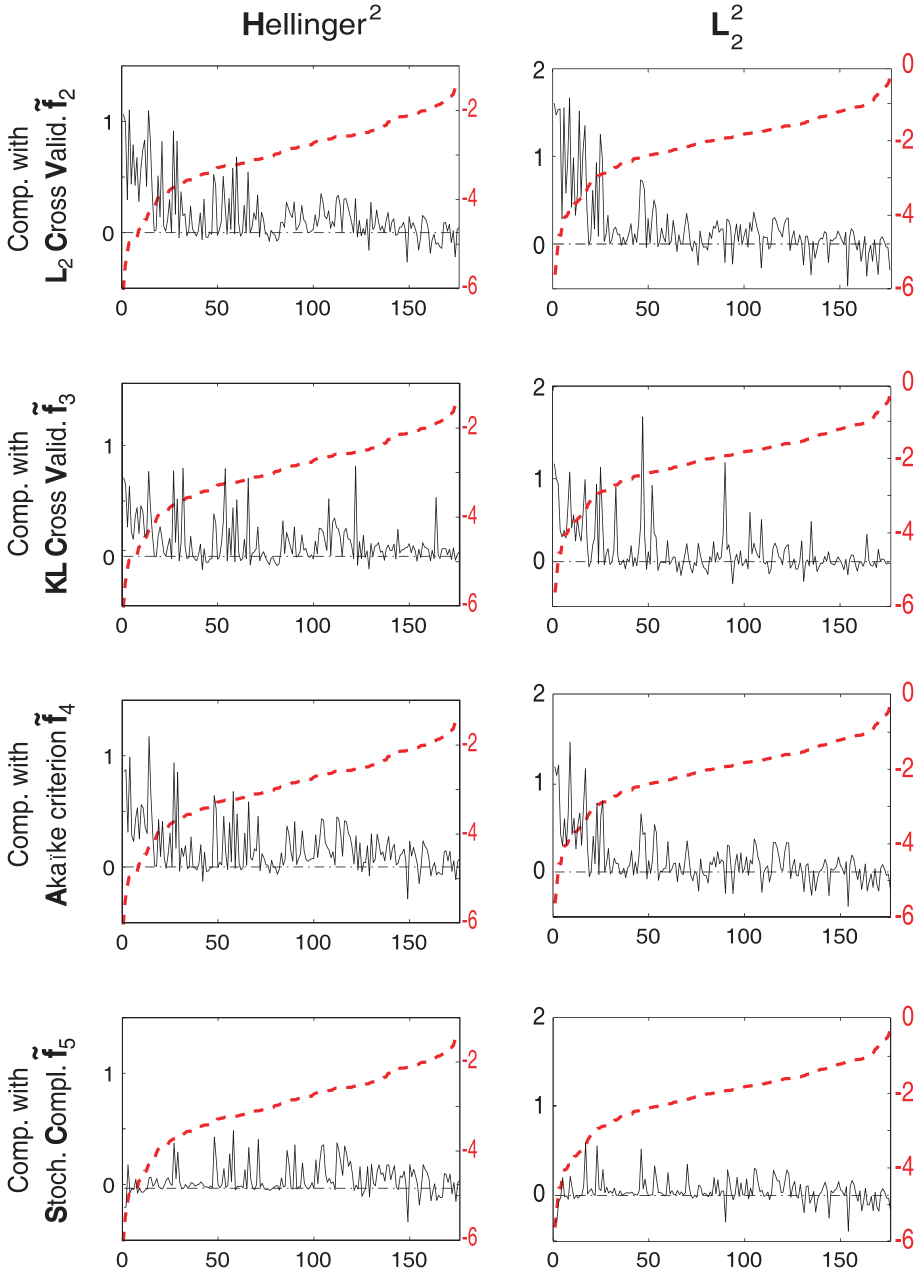

FIGURE 4. Comparison between $\tilde{f}_2$, $\tilde{f}_3$, $\tilde{f}_4$, $\tilde{f}_5$ and our method using Hellinger^2, $\mathbb{L}_1$, $\mathbb{L}_2^2$, $\mathbb{L}_5^5$ losses. The black line gives the binary logarithm of the normalized ratio between the risks of $\tilde{f}_j$ and $\tilde{f}_1$, for all $(f, n) \in \mathcal{F}$ sorted in increasing order of the binary logarithm of the normalized risks of the oracle (grey dashed line with scale reference on the right side).

different from our criteria. That all the methods ($\tilde{f}_7$ to $\tilde{f}_{10}$) which define an asymptotically optimal binwidth from a smoothness assumption on the underlying density do not work well for estimating discontinuous densities is natural as well. Since $\tilde{f}_{14}$ only chooses dyadic partitions, this tends to result in an increased bias.

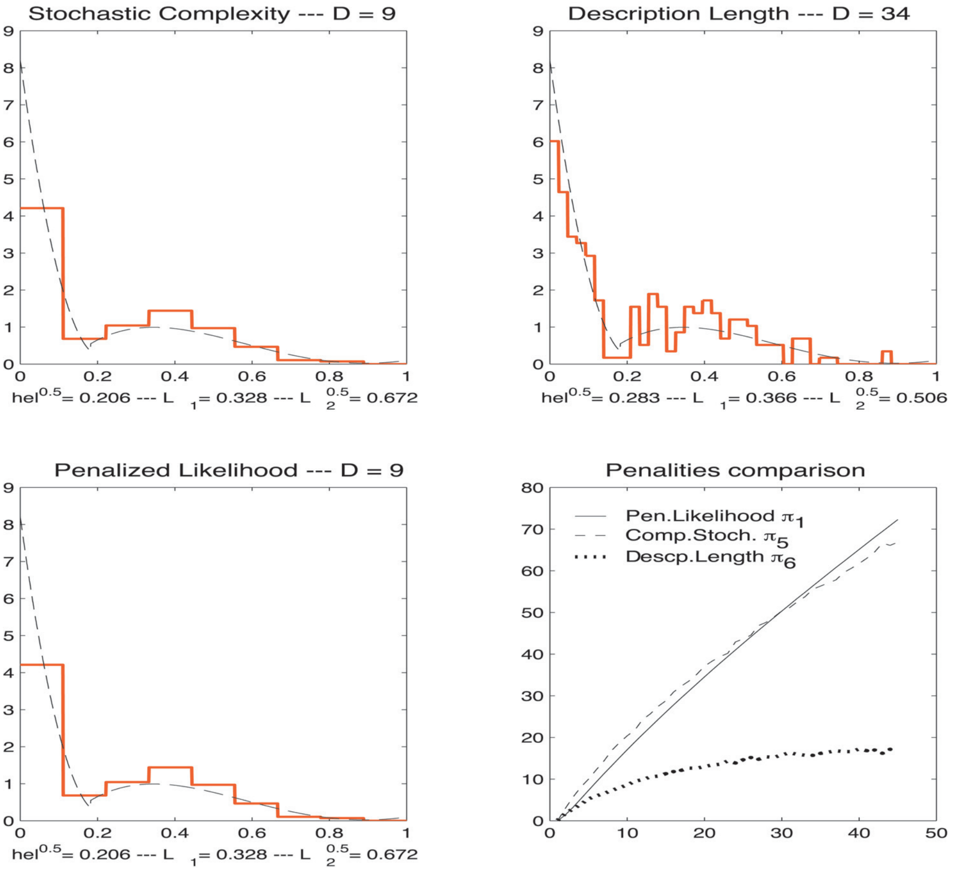

FIGURE 5. Stochastic complexity and Minimum Description Length (viewed as penalized likelihood methods) and our method – Comparison of penality terms.

Actually, all the methods that work reasonably well are either based on cross-validation or some complexity penalization arguments. It was therefore rather surprising for us to notice that the two estimators studied by Hall and Hannan [13], which are asymptotically equivalent and have similar performances for moderate sample sizes according to the authors, appear to behave quite differently in our study, the estimator based on stochastic complexity being much better than the one based on minimum description length. This is probably due to the fact that this equivalence is really of an asymptotic nature and that the testing densities in Hall an Hannan are very smooth (normal and beta) while ours are not. Rewriting the three estimators $\tilde{f}_j$ with $j = 1, 5, 6$ as $\hat{f}_{\hat{D}_j}$ where $\hat{D}_j$ is the maximizer of $L_n(D) - \pi_j(D)$, we compared the behaviours of π_1, π_5 and π_6 for different simulated examples. Note that here $\pi_1(D) = \mathrm{pen}(D)$ as defined by (1.2). The examples show that π_1 and π_5 are rather close while π_6 tends to be much smaller leading to larger values for $\hat{D}_6$. An illustration of the phenomenon is shown in Figure 5.

4.3. An additional comparison study

Since we partly used the same set of densities for both the optimization and comparison steps of our study, the results of the simulation study could be biased by the fact that we optimized our estimator for this special set of densities. A referee, that we thank for his many useful comments, suggested that we also test our method against the others on some very classical densities like the normal. He was also interested in seeing how our

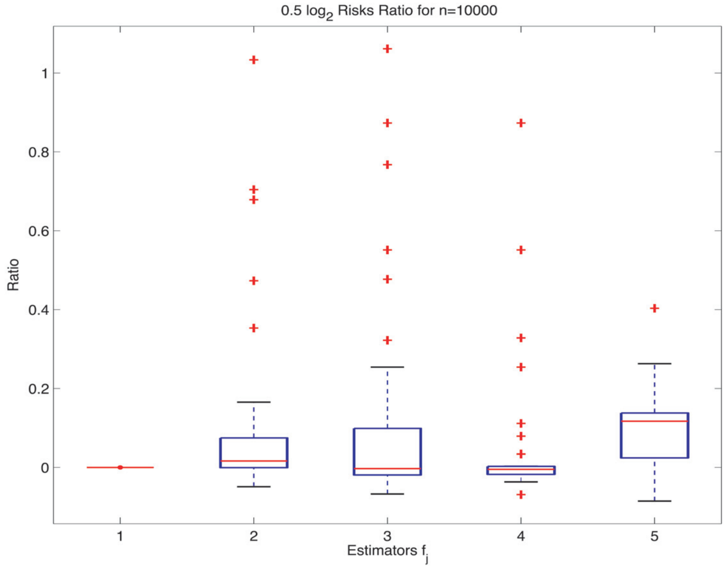

Figure 6. Boxplot of the binary logarithms of the normalized ratio between the risks of $\tilde{f}_2$ to $\tilde{f}_5$ and $\tilde{f}_1$ for $n = 10\,000$.

method would perform on larger samples. Since the computation of oracles is extremely time-consuming (their computation time is of order Cn^2, where n is the sample size and C a large constant), it was not realistic to compute oracles for samples of size 10 000 and in this part of our study, we contented ourselves to evaluate the risks of the 14 procedures and computed the binary logarithm of the ratios between the Hellinger risk of method j and ours, *i.e.* $\log_2\left[\overline{M}^{\star}_n(f,\tilde{f}_j,\ell_0)/\overline{M}^{\star}_n(f,\tilde{f}_1,\ell_0)\right]$.

4.3.1. *Larger samples*

We first conducted the study of Section 4.2 on our initial set $\mathcal{F}$ with samples of size 10 000 and computed the logarithm of the ratios between the risk of methods 2 to 5 and ours with Hellinger risk $\log_2[\overline{M}^{\star}_n(f,\tilde{f}_j,\ell_0)/\overline{M}^{\star}_n(f,\tilde{f}_1,\ell_0)]$. The resulting boxplots are given in Figure 6. Not surprisingly, AIC's based method $\tilde{f}_4$ and our method give very similar results for large samples since the corresponding penalties are asymptotically equivalent.

4.3.2. *Some classical densities*

We selected six smooth, but not necessarily compactly supported densities g_j, $1 \leq j \leq 6$ namely the standard normal g_1, two Gaussian mixtures g_2 and g_3, respectively $(1/4)\mathcal{N}(0,1)+(3/4)\mathcal{N}(2,1/16)$ and $(3/4)\mathcal{N}(0,1)+(1/4)\mathcal{N}(3,1/9)$, the exponential with parameter one g_4, the Student with 3 degrees of freedom g_5 and the uniform g_6. Then we tested the different methods on them. For each sample (apart from the uniform case), we estimated the support by the interval between the smallest and the largest observation, performed an affine transform to change it to $[0;1]$ and applied the various methods as usual, then going back to the initial scale to compute the loss, integrating it over the unbounded support of the true density, not only on the estimated support. We give in Figure 7 the plots of $\log_2[\overline{M}^{\star}_n(g_j,\tilde{f}_j,\ell_0)/\overline{M}^{\star}_n(g_j,\tilde{f}_1,\ell_0)]$ for each of our 6 test densities g_j and $n = 25, 100, 250, 1000$ and 10 000, each dotted line corresponding to a sample size (given on the left-hand side of the figures) and the "+" to the different methods in order of their numbers from 2 to 14. For each sample size, the baseline corresponds to 0, *i.e.* the performance of our estimator and the value of its risk is indicated on the right-hand side of the figure. The results confirm that our method performs quite well on those classical densities. Apart from the uniform distribution, which appears to be a special case, the results are relatively

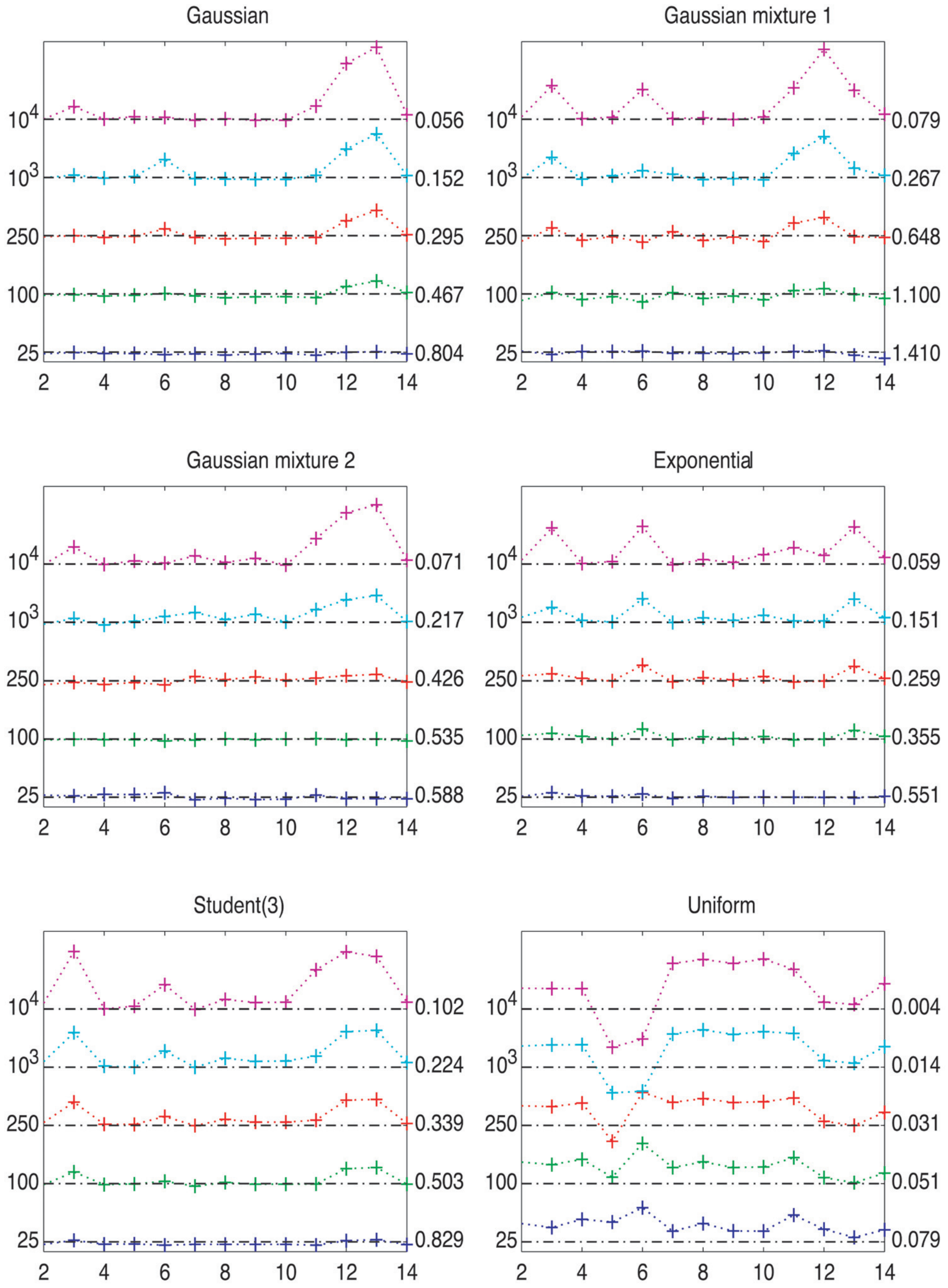

Figure 7. Comparison of the performances of the test estimators for various sample sizes and 6 classical densities.

homogeneous. It appears that $\tilde{f}_3$ does not behave so well here. The situation for the uniform is special. For small sample sizes and although not specially designed for this case which was not included in our test set, our method surprisingly outperforms all the others apart from $\tilde{f}_{13}$. For larger samples, the estimators based on

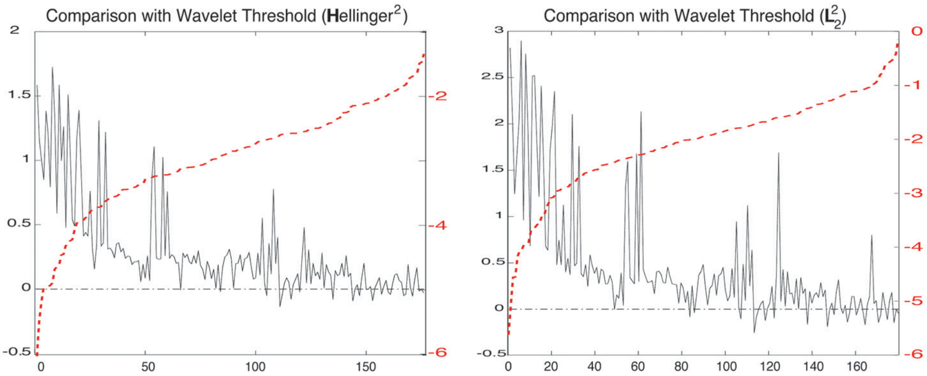

FIGURE 8. Comparison between the wavelet estimator $\tilde{f}_{15}$ and our method for Hellinger^2 and $\mathbb{L}_2^2$ losses. The black line gives the binary logarithms of the normalized ratio between the risks of $\tilde{f}_{15}$ and $\tilde{f}_1$, for all $(f, n) \in \mathcal{F}$ sorted in increasing order of the binary logarithm of the normalized risks of the oracle (represented in grey dashed line with scale reference on the right side).

stochastic complexity and minimum description length, which penalize more and tend to choose smaller values of $\hat{D}$, become substantially better than the others although our method still outperforms the remaining ones.

4.4. **About irregular histograms**

To conclude our study, let us illustrate the fact, mentioned in our introduction, that a model selection method based on irregular histograms, although intuitively more attractive, is not necessarily systematically better than ours because of its definitely higher complexity. Not only it looses the computational simplicity and speed connected with the use of regular histograms, but, more seriously, the bias reduction due to the use of a huge number of partitions instead of the much smaller set of regular ones, is often compensated by an increase of the stochastic error due to the difficulty of selecting a model among a large number of potential ones.

It has actually been shown (Birgé and Massart [4]) in a different, but related stochastic framework, namely variable selection for Gaussian regression, that penalization methods do require a heavier penalty when one selects among a very large family of models. In particular, complete variable selection (which is the analogue of selection among irregular partitions in our case) requires $\log n$ factors for the penalty which results in an increased value of the stochastic component of the risk, as compared to ordered variable selection (which is the analogue of selection among regular partitions).

To support these arguments, we added to our comparison study a recent method, $\tilde{f}_{15}$, based on wavelet thresholding, since such estimators are considered as very powerful and therefore rather fashionable nowadays. In order to have a fair comparison in terms of bias, we used the Haar wavelet basis to get histogram-like estimators, in connection with a method from Herrick, Nason and Silverman [15] which appeared to be the best among the different wavelet methods for density estimation we tried. Note that the resulting estimator is not a regular histogram since wavelet thresholding amounts to selecting a partition among irregular dyadic ones. The graphs of $\log_2 \left[\overline{M}_n^\star(f, \tilde{f}_{15}, \ell_i)/\overline{M}_n^\star(f, \tilde{f}_1, \ell_i)\right]$, based on the same set of densities and provided by Figure 8, show that this estimator, although selecting among many irregular partitions, does not, one the whole, perform better than the others.

5. Appendix

5.1. **Proof of Theorem 1**

Since the risk R_n of $\hat{f}_{\mathcal{I}}$ is given by

$$R_n = \mathbb{E}\left[h^2(f,\hat{f}_{\mathcal{I}})\right] = \sum_{j=1}^{D} \mathbb{E}\left[\frac{1}{2}\int_{I_j}\left(\sqrt{f(x)} - \sqrt{\frac{N_j}{n\mu(I_j)}}\right)^2 \mathrm{d}\mu(x)\right], \tag{5.1}$$

it suffices to bound each term in the sum. The generic term can be written, omitting the indices, setting $l = \mu(I)$, $p = \int_I f\,\mathrm{d}\mu$ and denoting by N a binomial $\mathcal{B}(n,p)$ random variable, as

$$\begin{aligned} R(I) &= \mathbb{E}\left[\frac{1}{2}\int_I\left(\sqrt{f(x)} - \sqrt{\frac{N}{nl}}\right)^2 \mathrm{d}\mu(x)\right] \\ &= \frac{1}{2}\left(\int_I f(x)\,\mathrm{d}\mu(x) + \mathbb{E}\left[\frac{N}{n}\right]\right) - \mathbb{E}\left[\sqrt{\frac{N}{n}}\right]\int_I\sqrt{\frac{f(x)}{l}}\,\mathrm{d}\mu(x) \\ &= p - \mathbb{E}\left[\sqrt{\frac{N}{n}}\right]\int_I\sqrt{\frac{f(x)}{l}}\,\mathrm{d}\mu(x). \end{aligned}$$

Introducing $\overline{f}\mathbb{1}_I = l^{-1}p\mathbb{1}_I$ and $h^2 = h^2(f\mathbb{1}_I, \overline{f}\mathbb{1}_I)$, we notice that

$$h^2 = \frac{1}{2}\int_I\left(\sqrt{f(x)} - \sqrt{\frac{p}{l}}\right)^2 \mathrm{d}\mu(x) = p - \sqrt{p}\int_I\sqrt{\frac{f(x)}{l}}\,\mathrm{d}\mu(x),$$

which implies that

$$R(I) = p - \left(p - h^2\right)\mathbb{E}\left[\sqrt{\frac{N}{np}}\right] = h^2\mathbb{E}\left[\sqrt{\frac{N}{np}}\right] + p\left(1 - \mathbb{E}\left[\sqrt{\frac{N}{np}}\right]\right).$$

The conclusion then follows from the next lemma and (5.1).

Lemma 1. *Let N be a binomial random variable with parameters n and p, $0 < p < 1$, then*

$$\mathbb{E}\left[\sqrt{\frac{N}{np}}\right] > 1 - \frac{1-p}{2np} \qquad \text{and} \qquad \mathbb{E}\left[\sqrt{\frac{N}{np}}\right] = 1 - \frac{1-p}{8np}\left[1 + \mathcal{O}\left(\frac{1}{np}\right)\right].$$

Proof. Setting $Z = N - np$, we write $\mathbb{E}\left[\sqrt{N/(np)}\right] = \mathbb{E}\left[\sqrt{1 + Z/(np)}\right]$. The first inequality follows from the fact that, for $u \geq -1$, $\sqrt{1+u} \geq 1 + u/2 - u^2/2$, $\mathbb{E}[Z] = 0$ and $\mathrm{Var}(Z) = np(1-p)$. To get the asymptotic result, we use the more precise inequality

$$1 + \frac{u}{2} - \frac{u^2}{8} + \frac{u^3}{16} - \frac{5u^4}{16} \leq \sqrt{1+u} \leq 1 + \frac{u}{2} - \frac{u^2}{8} + \frac{u^3}{16}$$

together with the moments of order three and four of Z:

$$\mathbb{E}\left[Z^3\right] = np(1-p)(1-2p); \quad \mathbb{E}\left[Z^4\right] = np(1-p)\left[1 - 6p + 6p^2 + 3np(1-p)\right]. \qquad \square$$

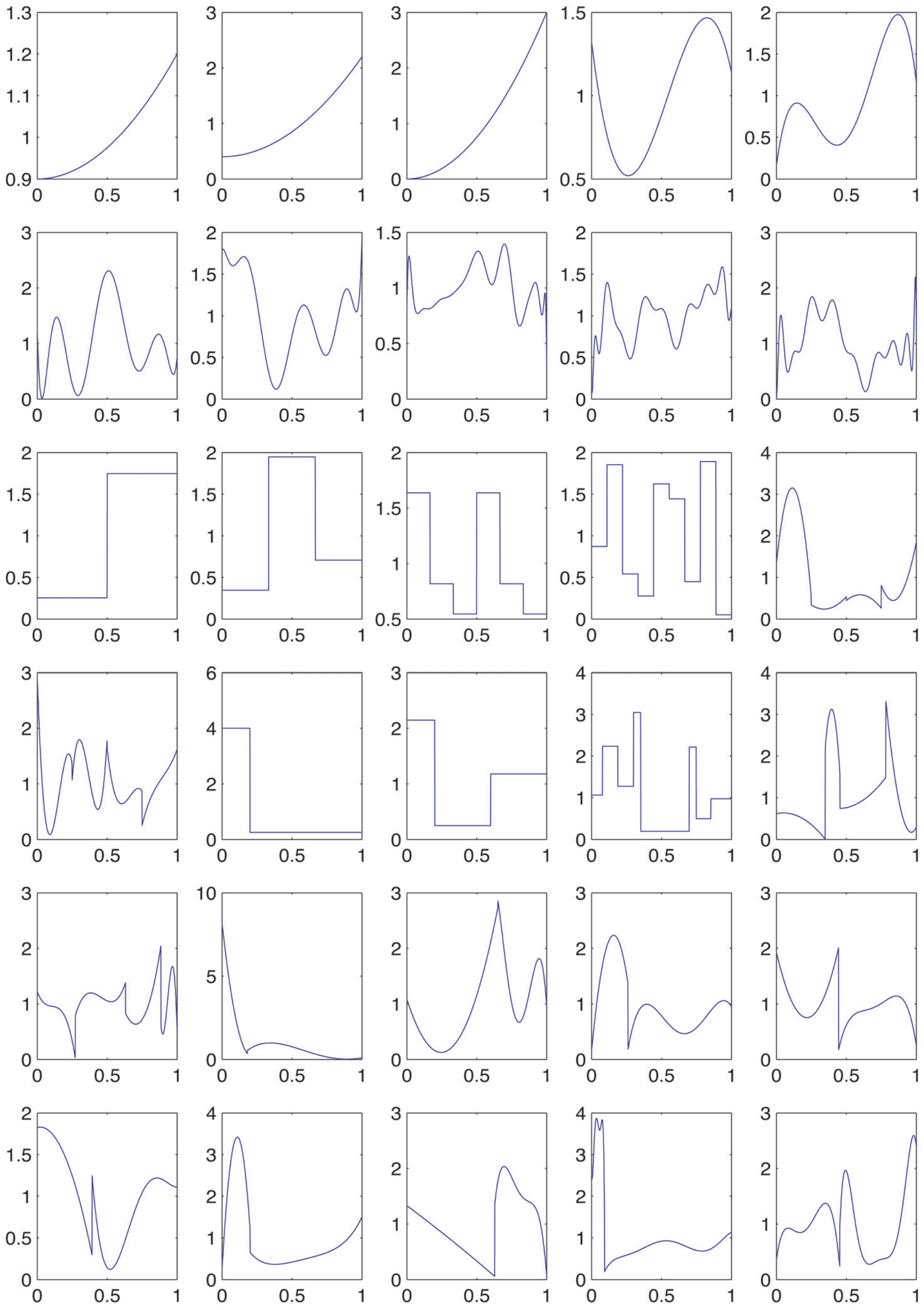

FIGURE 9. The used densities.

5.2. **Our set of test estimators**

Each of the estimators $\tilde{f}_k$, $1 \leq k \leq 14$, that we consider in our study (see Sect. 4.2) is based on a specific selection method, $\hat{D}_k(X_1, \ldots, X_n)$, which derives the number of bins from the data, resulting in $\tilde{f}_k = \hat{f}_{\hat{D}_k}$ with

$\hat{f}_D$ given by (2.1). For definiteness, we recall more precisely in this section the definitions of the various methods involved, adjusted to our particular situation of a support of length one. In the formulas below N_j, as defined in (1.1), denotes the number of observations falling in the jth bin.

The 6 first methods we considered are based on the maximization of some specific criterion with respect to the number D of bins. We recall that our estimator $\tilde{f}_1$ is based on the maximization of

$$\sum_{j=1}^{D} N_j \log N_j + n \log D - \left[D - 1 + (\log D)^{2.5}\right].$$

For $\mathbb{L}_2$ and Kullback cross-validation rules $\hat{D}_2$ and $\hat{D}_3$, the functions to maximize are given respectively (Rudemo [21], p. 69 and Hall [12] p. 452) by

$$\frac{D(n+1)}{n^2} \sum_{j=1}^{D} N_j^2 - 2D \qquad \text{and} \qquad \sum_{j=1}^{D} N_j \log(N_j - 1) + n \log D,$$

while AIC criterion $\hat{D}_4$ (Akaike [1]) corresponds to the maximization of $\sum_{j=1}^{D} N_j \log N_j + n \log D - (D - 1)$. The estimators $\tilde{f}_5$ and $\tilde{f}_6$, respectively based on stochastic complexity and minimum description length considerations, involve the maximization (Hall and Hannan [13]) of

$$D^n \frac{(D-1)!}{(D+n-1)!} \prod_{j=1}^{D} (N_j)!$$

and

$$\sum_{j=1}^{D} (N_j - 1/2) \log(N_j - 1/2) - (n - D/2) \log(n - D/2) + n \log D - (D/2) \log n.$$

Estimators $\tilde{f}_7$ to $\tilde{f}_{10}$ are all based on data driven evaluations $\hat{l}_k, 7 \leq k \leq 10$ of the binwidth. Since such evaluations do not lead to an integer number of bins when the support is $[0, 1]$, we took for $\hat{D}_k$ the integer which was closest to $\hat{l}_k^{-1}$. For $k = 7, 8, 9$, the respective suggestions for $\hat{l}_k$ by Taylor [24] or Kanazawa [17], Devroye and Györfi [9] and Scott [22] are

$$2.29\hat{\sigma}^{2/3} n^{-1/3}; \qquad 2.72\hat{\sigma} n^{-1/3}; \qquad \text{and} \qquad 3.49\hat{\sigma} n^{-1/3},$$

where $\hat{\sigma}^2$ denotes some estimator of the variance. We actually used for $\hat{\sigma}^2$ the unbiased version of the empirical variance. The previous binwidth estimates are actually based on the assumption that the shape of the underlying density is not far from a normal $\mathcal{N}(\mu, \sigma^2)$ distribution. There are various alternatives to this so-called Normal rule, including the oversmoothing method of Terrell [25] and a plug-in method of Wand [26], p. 62. We actually implemented Wand's method for the evaluation of $\hat{l}_{10}$, using the one-stage rule that he denotes by $\tilde{h}_1$ with $M = 400$ in his formula (4.1).

For $\tilde{f}_{11}$, we merely used Sturges'rule with $\hat{D}_{11}$ the integer closest to $1 + \log_2 n$. Daly [7] suggests to take $\hat{D}_{12}$ as the minimal value of D such that

$$(D+1) \sum_{j=1}^{D+1} N_j^2(D+1) - D \sum_{j=1}^{D} N_j^2(D) < \frac{n^2}{n+1},$$

where $N_j(D)$ denotes the number of observations falling in the jth bin of the regular partition with D bins. We implemented for $\tilde{f}_{13}$ the method given by He and Meeden [14], without the restriction they impose that the number of bins should be chosen between 5 and 20 since such a restriction leads to poor results for small sample sizes. It was replaced by the less restrictive $D > 1$. We also computed $\tilde{f}_{14}$ according to the method given in Chapter 10 of Devroye and Lugosi [10]. More precisely we used the histograms build by data splitting as described in their Section 10.3 with a maximal number of dyadic bins bounded by $2n$ (in order to avoid an algorithmic explosion) and a value of m set to the integer part of $n/2$. We actually also experimented smaller values of m but did not notice any improvement.

The last estimator $\tilde{f}_{15}$ we used for the comparison is not a regular histogram but a piecewise constant function derived from an expansion within the Haar wavelet basis, the construction following the recommendations of Herrick, Nason and Silverman [15] with the use of the normal approximation with p-value 0.01 and a finest resolution level set to $\log_2 U$ where U is the minimum of n^2 (to avoid an algorithmic explosion) and the inverse of the smallest distance between data.

Acknowledgements. We would like to express special thanks to the Associate Editor and referees for their many useful comments and suggestions.

References

[1] H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19** (1974) 716–723.

[2] A.R. Barron, L. Birgé and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113** (1999) 301–415.

[3] L. Birgé and P. Massart, From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen and G. Yang, Eds., Springer-Verlag, New York (1997) 55–87.

[4] L. Birgé and P. Massart, Gaussian model selection. *J. Eur. Math. Soc.* **3** (2001) 203–268.

[5] G. Castellan, *Modified Akaike's criterion for histogram density estimation.* Technical Report. Université Paris-Sud, Orsay (1999).

[6] G. Castellan, Sélection d'histogrammes à l'aide d'un critère de type Akaike. *CRAS* **330** (2000) 729–732.

[7] J. Daly, The construction of optimal histograms. *Commun. Stat., Theory Methods* **17** (1988) 2921–2931.

[8] L. Devroye, *A Course in Density Estimation.* Birkhäuser, Boston (1987).

[9] L. Devroye, and L. Györfi, *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York (1985).

[10] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York (2001).

[11] D. Freedman and P. Diaconis, On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **57** (1981) 453–476.

[12] P. Hall, Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Relat. Fields* **85** (1990) 449–467.

[13] P. Hall and E.J. Hannan, On stochastic complexity and nonparametric density estimation. *Biometrika* **75** (1988) 705–714.

[14] K. He and G. Meeden, Selecting the number of bins in a histogram: A decision theoretic approach. *J. Stat. Plann. Inference* **61** (1997) 49–59.

[15] D.R.M. Herrick, G.P. Nason and B.W. Silverman, Some new methods for wavelet density estimation. *Sankhya, Series A* **63** (2001) 394–411.

[16] M.C. Jones, On two recent papers of Y. Kanazawa. *Statist. Probab. Lett.* **24** (1995) 269–271.

[17] Y. Kanazawa, Hellinger distance and Akaike's information criterion for the histogram. *Statist. Probab. Lett.* **17** (1993) 293–298.

[18] L.M. Le Cam, *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York (1986).

[19] L.M. Le Cam and G.L. Yang, *Asymptotics in Statistics: Some Basic Concepts.* Second Edition. Springer-Verlag, New York (2000).

[20] J. Rissanen, Stochastic complexity and the MDL principle. *Econ. Rev.* **6** (1987) 85–102.

[21] M. Rudemo, Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** (1982) 65–78.

[22] D.W. Scott, On optimal and databased histograms. *Biometrika* **66** (1979) 605–610.

[23] H.A. Sturges, The choice of a class interval. *J. Am. Stat. Assoc.* **21** (1926) 65–66.

[24] C.C. Taylor, Akaike's information criterion and the histogram. *Biometrika.* **74** (1987) 636–639.

[25] G.R. Terrell, The maximal smoothing principle in density estimation. *J. Am. Stat. Assoc.* **85** (1990) 470–477.

[26] M.P. Wand, Data-based choice of histogram bin width. *Am. Statistician* **51** (1997) 59–64.