HOW MANY BINS SHOULD BE PUT IN A REGULAR HISTOGRAM

Lucien Birgé, Yves Rozenholc

How many bins should be put
in a regular histogram

L. BIRGÉ & Y. ROZENHOLC

AVRIL 2002

Prépublication n^o 721

L. Birgé : Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII, 4 place Jussieu, Case 188, F-75252 Paris Cedex 05.

Y. Rozenholc : Laboratoire de Probabilités et Modèles Aléatoires, CNRS-UMR 7599, Université Paris VI & Université Paris VII, 4 place Jussieu, Case 188, F-75252 Paris Cedex 05 & Université du Maine.

How many bins should be put in a regular histogram

Lucien Birgé
Université Paris VI and UMR CNRS 7599

Yves Rozenholc
Université du Maine and UMR CNRS 7599

April 2002

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 set of data 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 a fully automatic and simple method to choose the number of bins of the partition from the data. It is based on a nonasymptotic evaluation of the performances of penalized maximum likelihood estimators in some exponential families 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.

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 bandwiths 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.

---

[0] *AMS 1991 subject classifications.* Primary 62G05; secondary 62E25.

Key words and phrases. Regular histogram, density estimation, penalized maximum likelihood, model selection.

Unfortunately, when faced to such a question, the professional statistician has no really 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 simple and fully automatic method to choose the number of bins to be used for building a regular histogram from data. The procedure 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.

The limitation to densities with support on $[0, 1]$ may seem restrictive, but real data can always be considered as generated from a compactly supported density. If the support is known, one can assume, by a suitable transformation of the data, that it is $[0, 1]$. If it is unknown, one can merely replace it by the range of the data since we cannot get any information of what happens outside this range without extra assumptions.

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 a number 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), all the methods we know about 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 (1999). For the specific case of histogram estimators, their results have been substantially improved by Castellan (1999) 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 then 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

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

The resulting estimator $\hat{f}_{\hat{D}}$ will be denoted $\tilde{f}_1$ from now on. A few simulation results describing graphically the performances of $\tilde{f}_1$ are given in Figure 1 .

In order to test this new procedure, we conducted another simulation study involving a large family of densities, sample sizes ranging from 25 to 1000 and different loss functions, in order to compare $\tilde{f}_1$ with a number of existing methods. The conclusion of this large scale empirical study, which is given in Section 4, is that our method, on the whole, outperforms all the others, although one of them, namely the one based on Rissanen's minimum complexity ideas (Rissanen, 1987) and introduced, in the context of histogram estimation, by Hall and Hannan (1988), 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 $[0,1]$. 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

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

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 theoretic 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 $\hat{D}(X_1, \ldots, X_n)$. The quality of such a procedure at f can be measured by the value of the ratio

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

where $\hat{f}_{\hat{D}}$ denotes the histogram estimator based on the regular partition with $\hat{D}$ pieces.

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 since, when $f = \mathbb{1}_{[0,1]}$, $D^{opt} = 1$,

$\hat{f}_1 = f$ and the ratio (2.3) is infinite. Even if f is different from $\mathbb{1}_{[0,1]}$ but is very close to it, $R_n(f, \hat{f}_1, \ell)$ may be arbitrarily close to zero and there is not hope to get a small value for (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 (2001). This implies that, in order to judge the quality of a procedure $\hat{D}$, we should only consider the ratio (2.3) for densities which are far enough — see the precise condition (3.1) below — from the uniform.

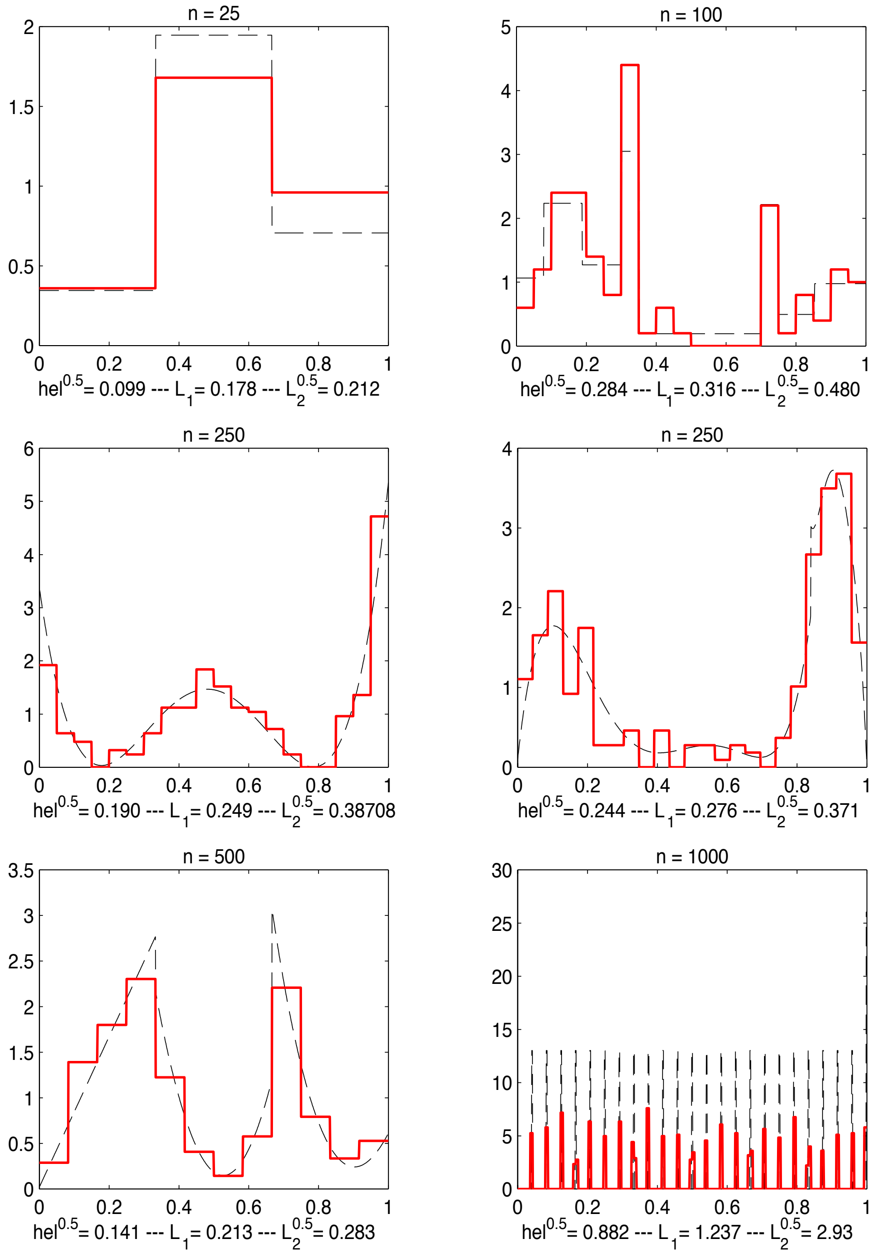

Figure 1: 6 examples, the dashed black line represents the true density and the thick grey line the estimator

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

$$h^2(f,g) = \frac{1}{2}\int_0^1 \left(\sqrt{f(y)} - \sqrt{g(y)}\right)^2 dy, \tag{2.4}$$

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

$$K(f,g) = \int_0^1 \log\left(\frac{f(y)}{g(y)}\right) f(y)\,dy \leq +\infty. \tag{2.5}$$

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 some continuity argument, 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,

$$\mathbb{E}_f\left[\left\|f-\hat{f}_D\right\|^2\right] = \mathbb{E}_f\left[\left\|\bar{f}_D-\hat{f}_D\right\|^2\right] + \left\|f-\bar{f}_D\right\|^2, \tag{2.6}$$

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 set of 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 (1979), Freedman and Diaconis (1981), Daly (1988), Wand (1997) and Birgé and Massart (1997).

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 (1985, p. 2) and the preface of Devroye (1987).

Although it is less popular, may be 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, 1986 or Le Cam and Yang, 2000). Finally, the results of Castellan (1999) that we use here are based on it.

Of course, the choice of a "nice" loss function is, in 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 (1993) 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 (1995). 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 since the ratio (2.3) involves it. There is actually nothing special with regular histograms from this point of view and a general result in this direction is as follows. Given an histogram estimator $\hat{f}_{\mathcal{I}}$ of the form (2.7) 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 (1999) but we were unable to find a non-asymptotic bound for it in the literature. It can actually be proved (see our Appendix) that

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.*

$$\hat{f}_{\mathcal{I}}(X_1,\ldots,X_n)=\frac{1}{n}\sum_{j=1}^{D}\frac{N_j}{\mu(I_j)}1\!\!1_{I_j},\quad \text{with } N_j=\sum_{i=1}^{n}1\!\!1_{I_j}(X_i). \tag{2.7}$$

Setting $p_j=\int_{I_j}fd\mu$, we get

$$\mathbb{E}\left[h^2(f,\hat{f}_{\mathcal{I}})\right]\le h^2(f,\bar{f}_{\mathcal{I}})+\frac{D-1}{2n}\quad \text{with } \bar{f}_{\mathcal{I}}=\sum_{j=1}^{D}\frac{p_j}{\mu(I_j)}1\!\!1_{I_j}. \tag{2.8}$$

Moreover

$$\mathbb{E}\left[h^2(f,\hat{f}_{\mathcal{I}})\right]=h^2(f,\bar{f}_{\mathcal{I}})+\frac{D-1}{8n}[1+o(1)], \tag{2.9}$$

when $n(\inf_{1\le j\le D}p_j)$ 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 (2.6).

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 (1999). 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

$$\mathrm{pen}(D)=c_1(D-1)\left(1+\sqrt{c_2L_D}\right)^2\quad \text{with } c_1>1/2,\quad L_D>0, \tag{2.10}$$

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

$$\sum_{D\ge 1}\exp[-(D-1)L_D]=\Sigma<+\infty. \tag{2.11}$$

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

$$\mathbb{E}_f \left[h^2(f, \hat{f}_{\hat{D}}) \right] \leq \kappa(c_1) \inf_{1 \leq D \leq \bar{D}} \left\{ K(f, \bar{f}_D) + n^{-1} \mathrm{pen}(D) \right\} + n^{-1} \kappa', \tag{2.12}$$

where κ, κ' are positive constants, κ depending on c_1, κ' on the parameters involved in the assumptions and in particular being increasing with respect to Σ. This bound and (2.11) 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 (2.10) 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, 1998, Lemma 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

$$\mathbb{E}_f \left[h^2(f, \hat{f}_{\hat{D}}) \right] \leq \kappa_1 \inf_{1 \leq D \leq \bar{D}} \left\{ h^2(f, \bar{f}_D) + \frac{D-1}{n} \right\} + \frac{\kappa_2}{n}, \tag{2.13}$$

where κ_2 is an increasing function of Σ. In view of Theorem 1, one can finally derive from (2.13) that, under suitable restrictions on f and for n large enough,

$$\mathbb{E}_f \left[h^2(f, \hat{f}_{\hat{D}}) \right] \leq \kappa_1' \mathbb{E}_f \left[h^2(f, \hat{f}_{D^{opt}}) \right] + \kappa_2'/n. \tag{2.14}$$

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 κ_2/n of moderate size when n is not very large. In order to derive a form of the 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 (2.9) 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 $\mathbb{E}_f \left[h^2(f, \hat{f}_D) \right] \gtrsim (8n)^{-1}$ for $D \geq 2$. Together with (2.14), 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 $\bar{f}_1 = 1\!\!1_{[0,1]}$, whatever f, which means that f is not too close to the uniform. This restriction confirms the arguments of Section 2.1.

If $c_1 = 1$, (2.10) can be written

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

Moreover, the constant c_2 is only known approximately and (2.11) requires that $(D-1)L_D$ tends to infinity with D. Since L_D should not be large because it influences the risk as shown by (2.12), it seems natural to look for penalties of the form $\mathrm{pen}(D) = D - 1 + g(D)$ where the function g tends to infinity but not too fast. We actually restricted our search to functions $g(x)$ of the three following types:

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

with $\alpha > 0$, varying the values of both parameters. We also replaced the restriction $1 \le D \le \bar{D}$ 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 practise.

3.2 The operational procedure

We proceeded in two steps. The first one was an *optimization step* to choose a convenient form for the function g just mentioned; 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. The chosen densities are of piecewise polynomial form. To define them we used the trivial partition with a single element which leads to continuous densities, and 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 also added to the resulting family some special piecewise constant densities which were known to be difficult to estimate and ended up with a set of 45 different densities, ranging from smooth to rather erratic, which are described in Figure 6 in the Appendix.

Many of these densities are far from 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.

We then selected a large range of values for n, namely $n = 25, 50, 100, 250, 500$ and 1000 and this resulted in a set $\mathcal{F}$ of 264 pairs (f, n) after we excluded, in view of the previous arguments, six pairs which did not satisfy the requirement (3.1).

In both steps, 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 $X_1^j, \ldots, X_n^j$, $1 \le j \le 1000$ 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: $\mathbb{P}_f[\ell(f, \tilde{f}(X_1, \ldots, X_n)) \le t]$ by their empirical counterparts

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

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 of the variables $\ell(f, \tilde{f}(X_1^j, \ldots, X_n^j))$. 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/histograms`.

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 $\mathcal{F}$. As we previously mentioned, the performance of a selection procedure $\hat{D}(g)$ based on the penalty involving some function g can be evaluated by a comparison of its risk with the optimal risk corresponding to $D = D^{opt}$.

Ideally, one would like to minimize the ratio

$$\overline{M}_n(f, \hat{f}_{\hat{D}(g)}, h^2) = \frac{\overline{R}_n(f, \hat{f}_{\hat{D}(g)}, h^2)}{\inf_{D \geq 1} \overline{R}_n(f, \hat{f}_D, h^2)}, \tag{3.3}$$

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 $D^{opt}(f, n)$ and we actually tried to minimize approximately with respect to g

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

for all values of k simultaneously. Again, this is not a well-defined problem and some compromises were needed, but we finally concluded that the choice $g(x) = (\log x)^{2.5}$ was the most satisfactory and ended up with the estimator $\tilde{f}_1$ described in the introduction.

3.4 Some 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,

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

with

$$p_0 = p_2 = 2, \qquad p_1 = 1 \qquad \text{and} \qquad p_3 = 5.$$

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 45 densities we started with, but some (at most three when $n = 25$), which are too close to the uniform, had to be excluded for smaller values of n 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 $n \geq 500$.

$i \setminus n$	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

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

Apart from these worst-case results, it is also 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 can "beat" the oracle. This is not so surprising since the oracle has a fixed number $D^{opt}(f,n)$ of bins (the one which minimizes the risk, i.e. the average loss) independently of the sample, while our estimator tries to optimize the number of bins for each sample and can therefore adjust to the peculiarities of the sample. This means that it is the very notion of an oracle which is questionable as a reference.

4 Comparison with previous methods

4.1 Some historical remarks

The first methods used to decide about the number of bins were just rules of thumbs and date back to Sturges (1926). According to Wand (1997) 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 (1979) and Freedman and Diaconis (1981) for the squared $\mathbb{L}_2$-loss, by Devroye and Györfi (1985) for the $\mathbb{L}_1$-loss and by Kanazawa (1993) 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. Since an estimation of c involves complicated computations, most authors suggest a rule of thumbs to evaluate it, typically: pretend that the true density is normal. Wand (1997) proposes to estimate c by kernel methods.

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 (1982). Theoretical comparisons between Kullback cross-validation and the AIC criterion are to be found in Hall (1990).

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 1974). Akaike's method is merely a penalized maximum likelihood method with penalty $\mathrm{pen}(D) = D - 1$ 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 (1987) 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 (1993). Related methods are those based on minimum description length and stochastic complexity due to Rissanen (see for instance Rissanen, 1987). Their specific implementation for histograms has been discussed in Hall and Hannan (1988).

Somewhat more exotic methods have been proposed by Daly (1988) and He and Meeden (1997), the second one being based on Bayesian bootstrap.

4.2 The comparison study

In order to evaluate the performances of our method and to compare it to previous ones, we selected 14 different estimators $\tilde{f}_2, \ldots, \tilde{f}_{15}$. The selection was based on the previous historical review and the precise description of the estimators is given 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 two estimators $\tilde{f}_{14}$ and $\tilde{f}_{15}$ which do not look for the optimal regular partition. $\tilde{f}_{14}$ is based on wavelet thresholding since such estimators are quite fashionable nowadays and also considered as very powerful. In order to have a fair comparison in terms of bias, we used the Haar wavelet basis in connection with a method from Herrick, Nason and Silverman (2001) which appeared to be the best among the different wavelet methods for density estimation we tried. Note that the resulting estimator, although piecewise constant, is not a regular histogram. The estimator $\tilde{f}_{15}$ is due to Devroye and Lugosi and described in Section 10.3 of Devroye and Lugosi (2001) but it selects only dyadic partitions.

We actually also studied the performances of modified versions of some of those estimates, as described by Rudemo (1982), Hall and Hannan (1988), Wand (1997) and various thresholding strategies for Haar wavelets. Since the performances of the modified methods were similar to or worse than those of the original estimators, we do not include them here.

4.2.1 Analysis of the results

To compare the performances of the various estimators, we computed, for all 264 pairs $(f, n) \in \mathcal{F}$, all estimators $\tilde{f}_k$, $1 \leq k \leq 15$ and the four selected loss functions, the values of $\overline{M}^{\star}_n(f, \tilde{f}_k, \ell_i)$.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. Dividing by the risk of the oracle does not influence the comparisons between the various estimators but substantially reduces the range of $\overline{Q}_{(0.95)}(n, f, \tilde{f}, \ell_i)$ when f varies in our test set which drastically improves the legibility of the results.

This simulation study 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 $f \in \mathcal{F}_n$. Our comparison of the estimates is 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 additional points give the values which are outside this range. Figure 2 shows the boxplots corresponding to all methods for $n = 25, 100$ and 1000, squared Hellinger and $\mathbb{L}_2$ losses. It is readily visible from these plots that estimators $\tilde{f}_6$ to $\tilde{f}_{15}$ are not satisfactory for $n \geq 100$ as compared to the others. The complete set of results shows that their performances for other values of n and loss functions are not better. This is why we do not include them in the final figures to improve legibility.

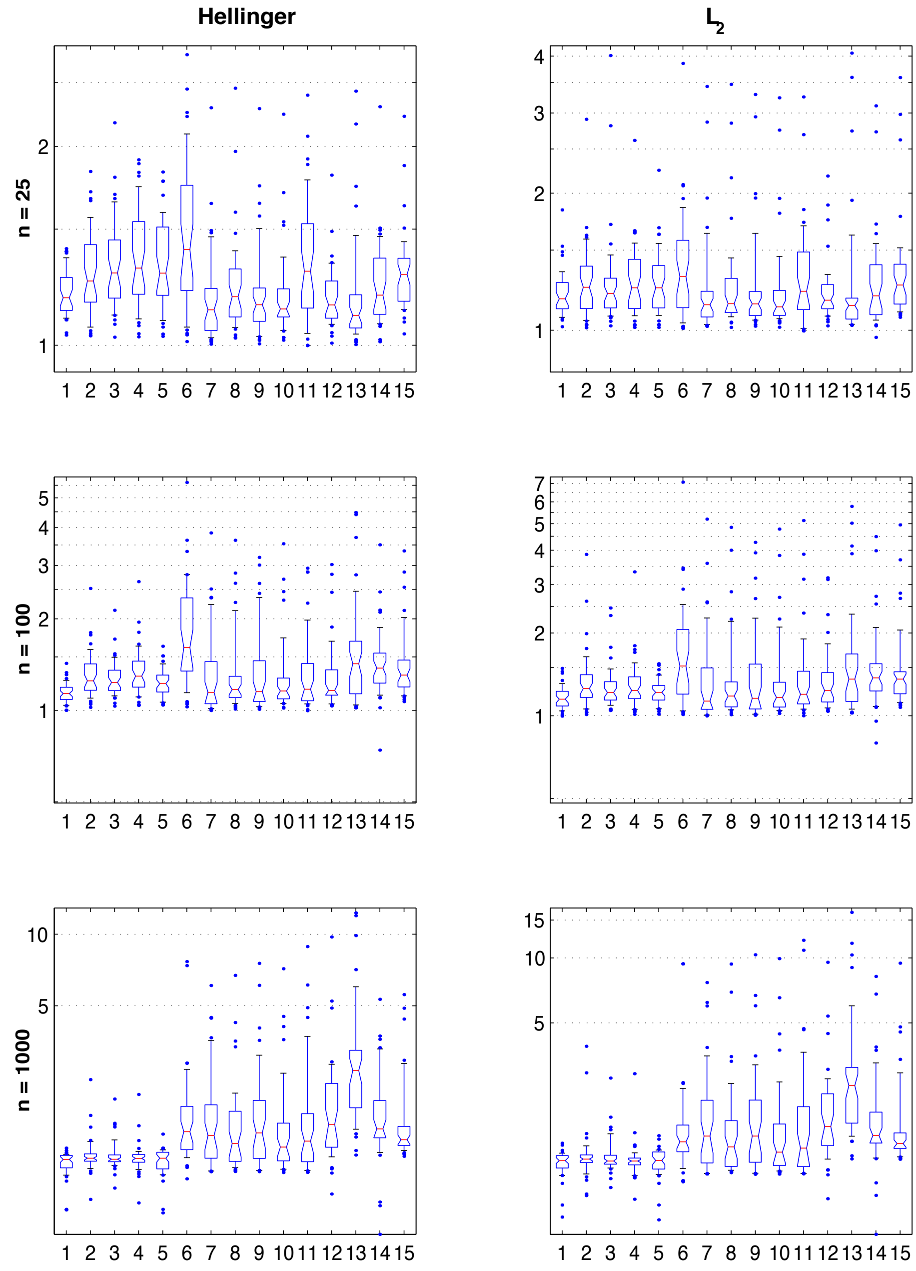

Figure 2: Hellinger (left) and $\mathbb{L}_2$ (right) mean ratios for $n = 25, 100, 1000$.

For the remaining 5 estimators, we provide hereafter in Figures 3 and 4 the complete series of boxplots corresponding to all values of n and ℓ. Clearly, our method is the best from this point of view, especially for the smaller values of n. Apart from ours, the best method seems to be $\tilde{f}_5$, based on stochastic complexity. If the performances of $\tilde{f}_5$ are roughly equivalent to those of $\tilde{f}_1$ for large values of n ($n \geq 250$), this is clearly not true for small n.

4.2.2 A few comments

As we previously noticed, some methods, corresponding to estimators $\tilde{f}_j$ with $6 \leq j \leq 15$ 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 theoretic arguments used by He and Meeden involve a very special loss function 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 either. Since $\tilde{f}_{15}$ only chooses dyadic partitions, this tend to result in an increased bias. Finally the method $\tilde{f}_{14}$, based on Haar waveled thresholding suffers from the same problem and moreover considers many more partitions than we do. It should rather be compared with selection procedures involving irregular partitions, since it is known (see for instance Birgé and Massart, 1997) that thresholding methods are equivalent to penalized methods for irregular dyadic partitions which do require heavier penalties.

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 (1988), 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 the 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

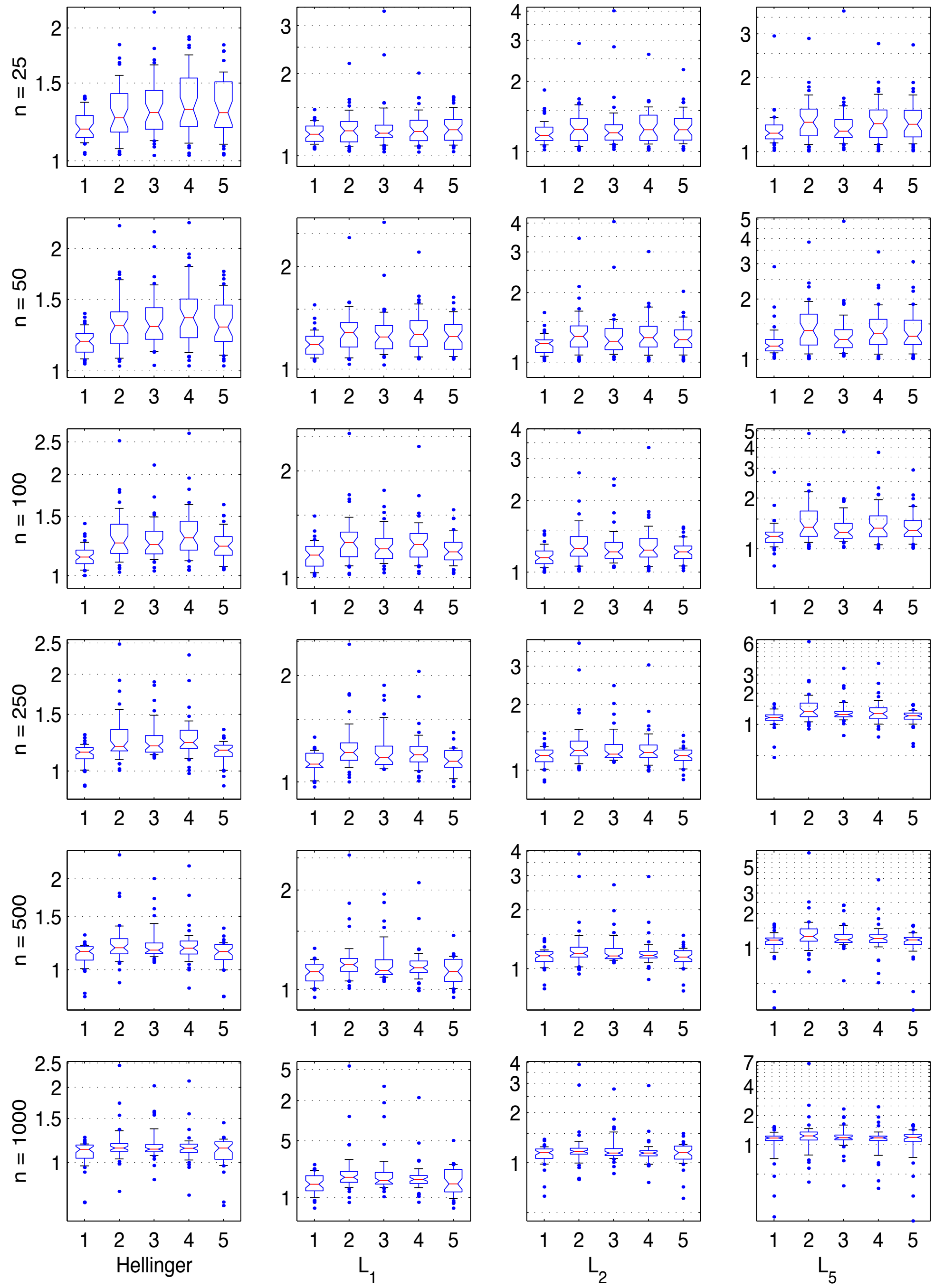

Figure 3: Hellinger, $\mathbb{L}_1$, $\mathbb{L}_2$ and $\mathbb{L}_5$ mean ratios for $n = 25, 50, 100, 250, 500, 1000$

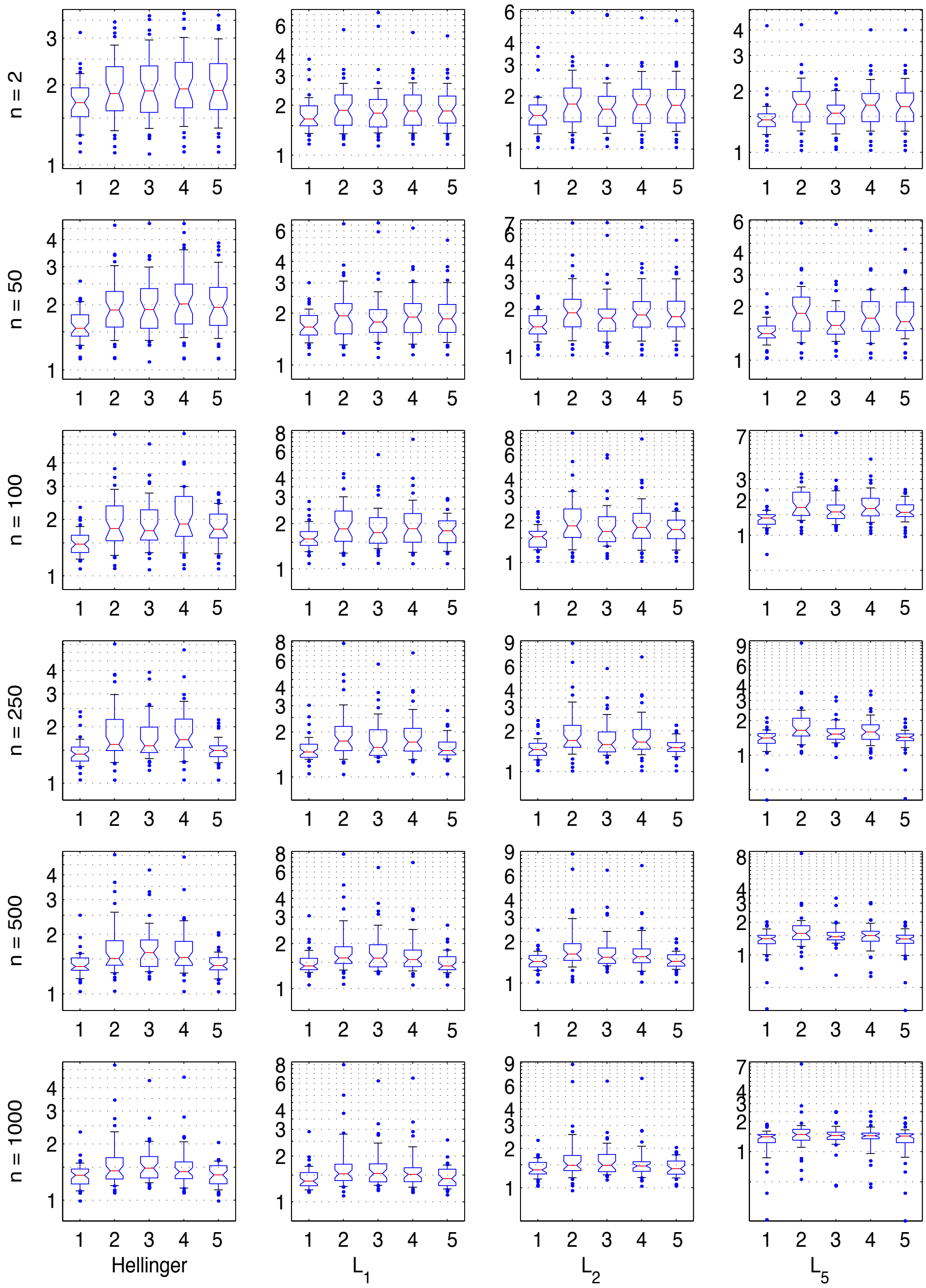

Figure 4: Hellinger, $\mathbb{L}_1$, $\mathbb{L}_2$ and $\mathbb{L}_5$ quantile ratios for $n = 25, 50, 100, 250, 500, 1000$

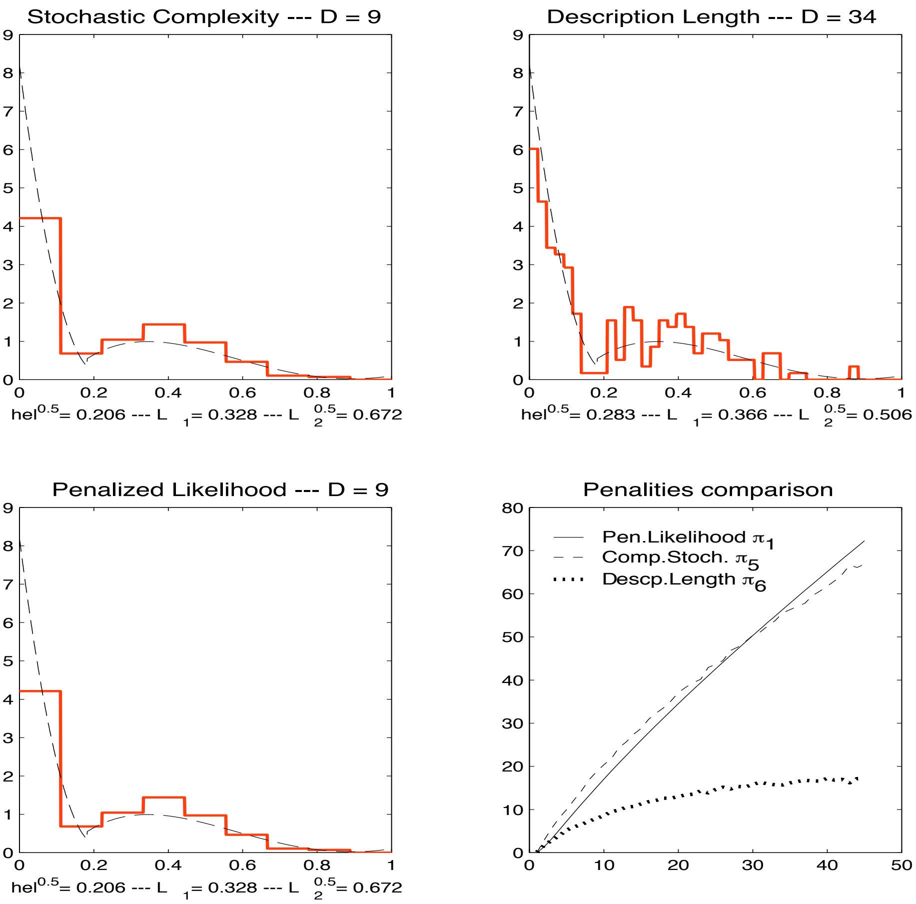

Figure 5: Stochastic complexity, Minimum Description Length viewed as penalized likelihoods and our method - Comparison of penality terms

5 Appendix

5.1 Proof of Theorem 1

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

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

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

$$R(I) \quad = \quad \mathbb{E}\left[\frac{1}{2}\int_I \left(\sqrt{f(x)} - \sqrt{\frac{N}{nl}}\right)^2 d\mu(x)\right]$$

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

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

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

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*

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

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:

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

5.2 Our set of test estimators

Apart from $\tilde{f}_{14}$, each of the estimators $\tilde{f}_k$, $1 \leq k \leq 15$, that we consider in our study (see Section 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 j-th bin.

The first 6 methods we considered are based on the maximization with respect to the number D of bins of some specific criterion. We recall that our estimator $\tilde{f}_1$ is based on the minimization 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, 1982 p. 69 and Hall, 1990 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, 1974) 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, 1988) 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 (1987) or Kanazawa (1993), Devroye and Györfi (1985) and Scott (1979) 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. To avoid the use of such a rule of thumbs, Wand (1997, p. 62) suggests a more complicated evaluation for $\hat{l}_{10}$ and we used the one-stage rule that he denotes by $\hat{h}_1$ 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 (1988) 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 $N_j(k)$ denotes the number of observations falling in the j-th bin of the regular partition with k bins. We implemented for $\tilde{f}_{13}$ the method given by He and Meeden (1997), 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 $\tilde{f}_{15}$ according to the method given in Chapter 10 of Devroye and Lugosi (2001). 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 an improvement.

The last estimator $\tilde{f}_{14}$ we used for the comparison is not a 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 (2001) 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.

References

AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. on Automatic Control* **19**, 716-723.

BARRON, A.R., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-415.

BIRGÉ, L. and MASSART, P. (1997). 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.), 55-87. Springer-Verlag, New York.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203-268.

CASTELLAN, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report. Universit é Paris-Sud, Orsay.

DALY, J. (1988). The construction of optimal histograms. *Commun. Stat., Theory Methods* **17**, 2921-2931.

DEVROYE, L. (1987). *A Course in Density Estimation.* Birkh?user, Boston.

DEVROYE, L. and GYÖRFI, L. (1985). *Nonparametric Density Estimation: The L_1 View.* John Wiley, New York.

DEVROYE, L. and LUGOSI, G. (2001). *Combinatorial Methods in Density Estimation.* Springer-Verlag, New York.

FREEDMAN, D. and DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheor. Verw. Geb.* **57**, 453-476.

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

HALL, P. and HANNAN, E.J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika* **75**, 705-714.

HE, K. and MEEDEN, G. (1997). Selecting the number of bins in a histogram: A decision theoretic approach. *J. Stat. Plann. Inference* **61**, 49-59.

HERRICK, D.R.M., NASON, G.P. and SILVERMAN, B.W. (2001). Some new methods for wavelet density estimation. Technical report.
`http://www.stats.bris.ac.uk/pub/ResRept/2001.html`

JONES, M.C. (1995). On two recent papers of Y. Kanazawa. *Statist. Probab. Lett.* **24**, 269-271.

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

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

Le CAM, L.M. and YANG, G.L. (2000). *Asymptotics in Statistics: Some Basic Concepts. Second Edition.* Springer-Verlag, New York.

RISSANEN, J. (1987). Stochastic complexity and the MDL principle. *Econ. Rev.* **6**, 85-102.

RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9**, 65-78.

SCOTT, D. W. (1979). Averaged shifted histograms: Effective nonparametric density estimators in several dimensions. *Ann. Statist.* **13**, 1024-1040.

STURGES, H. A. (1926). The choice of a class interval. *J. Am. Stat. Assoc.* **21**, 65-66.

TAYLOR, C. C. (1987). Akaike's information criterion and the histogram. *Biometrika.* **74**, 636-639.

WAND, M. P. (1997). Data-based choice of histogram bin width. *J. Am. Stat. Assoc.* **51**, 59-64

Lucien BIRGÉ
UMR 7599 "Probabilités et modèles aléatoires"
Laboratoire de Probabilités, boîte 188
Université Paris VI, 4 Place Jussieu

F-75252 Paris Cedex 05
France
e-mail: lb@ccr.jussieu.fr

Yves ROZENHOLC
UMR 7599 "Probabilités et modèles aléatoires"
Laboratoire de Probabilités, boîte 188
Université Paris VI, 4 Place Jussieu
F-75252 Paris Cedex 05
France
e-mail: yves.rozenholc@math.jussieu.fr

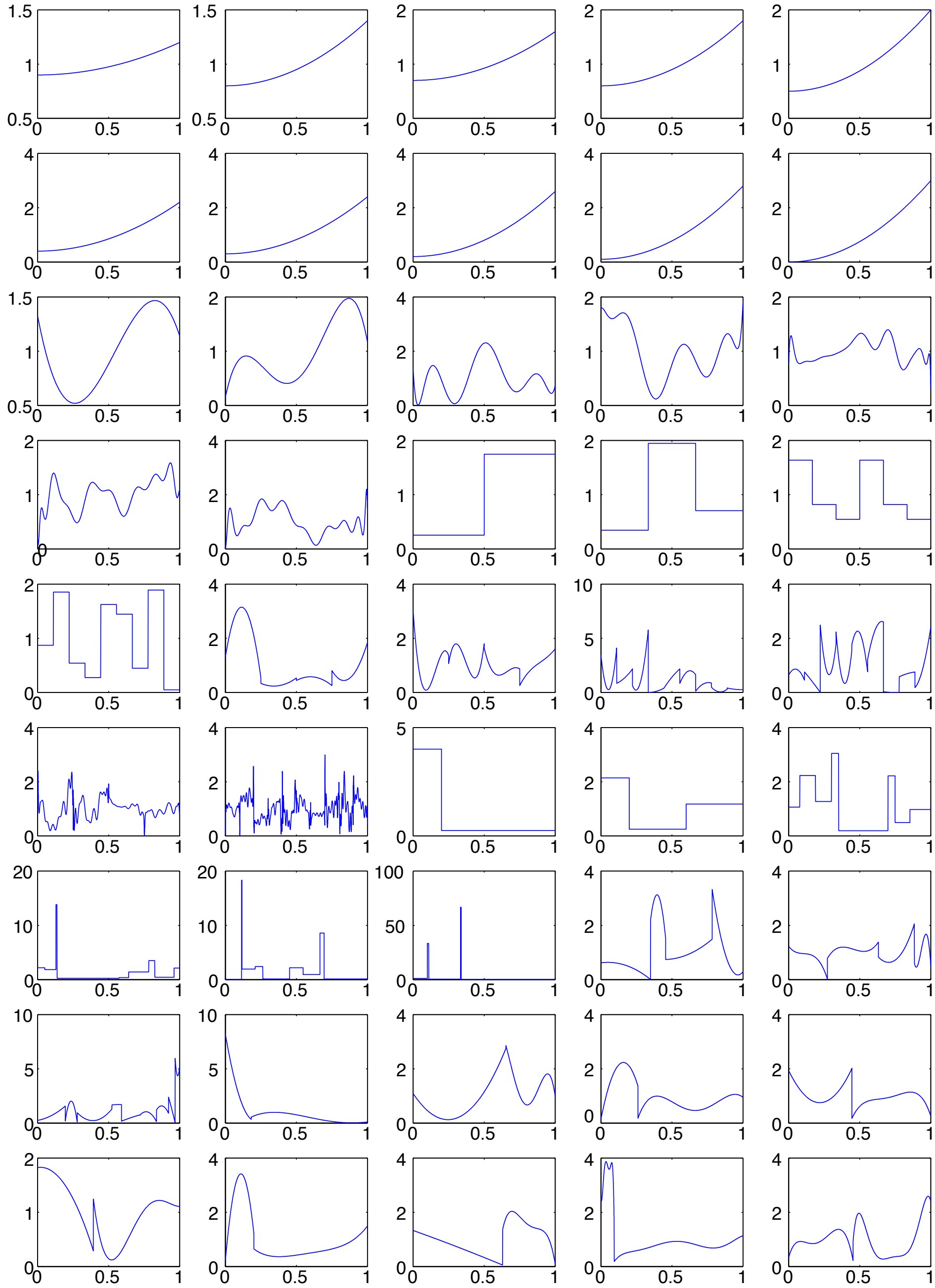

Figure 6: The used densities