Combining regular and irregular histograms by penalized likelihood

Yves Rozenholc, Thoralf Mildenberger, Ursula Gather

Combining Regular and Irregular Histograms by Penalized Likelihood

Yves Rozenholc[a], Thoralf Mildenberger[*,b], Ursula Gather[b]

[a] *UFR de Mathématiques et d'Informatique, Université Paris Descartes, MAP5 - UMR CNRS 8145, 45, Rue des Saints-Pères, 75270 Paris CEDEX, France*
[b] *Fakultät Statistik, Technische Universität Dortmund, 44221 Dortmund, Germany*

Abstract

A new fully automatic procedure for the construction of histograms is proposed. It consists of constructing both a regular and an irregular histogram and then choosing between the two. To choose the number of bins in the irregular histogram, two different penalties motivated by recent work in model selection are proposed. A description of the algorithm and a proper tuning of the penalties is given. Finally, different versions of the procedure are compared to other existing proposals for a wide range of densities and sample sizes. In the simulations, the squared Hellinger risk of the new procedure is always at most twice as large as the risk of the best of the other methods. The procedure is implemented in an R-Package.

Key words: irregular histogram, density estimation, penalized likelihood, dynamic programming

[*]Corresponding author. Tel.: +49 231 755 3129; fax: +49 231 755 5305; web: http://www.statistik.tu-dortmund.de/mildenberger_en.html

Email addresses: yves.rozenholc@parisdescartes.fr (Yves Rozenholc), mildenbe@statistik.tu-dortmund.de (Thoralf Mildenberger), gather@statistik.tu-dortmund.de (Ursula Gather)

1. Introduction

For a sample $(X_1, X_2, \ldots, X_n)$ of a real random variable X with an unknown density f w.r.t. Lebesgue measure, we denote the realizations by $(x_1, x_2, \ldots, x_n)$ and the realizations of the order statistics by $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. The goal in nonparametric density estimation is to construct an estimate $\hat{f}$ of f from the sample. In this work, we focus on estimation by histograms, which are defined as piecewise constant densities. The procedure we propose consists of constructing both a regular and an irregular histogram (both to be defined below) and then choosing between the two. Although other types of nonparametric density estimators are known to be superior to histograms according to several optimality criteria, histograms still play an important role in practice. The main reason is their simplicity and hence their interpretability (Birgé and Rozenholc, 2006). Often, the histogram is the only density estimator taught to future researchers in non-mathematical subject areas, usually introduced in an exploratory context without reference to optimality criteria.

We first introduce histograms and describe the connection to Maximum Likelihood estimation: Given $(x_1, x_2, \ldots, x_n)$ and a set of densities $\mathcal{F}$, the maximum likelihood estimate – if it exists – is given by an element $\hat{f} \in \mathcal{F}$ that maximizes the likelihood $\prod_{i=1}^{n} f(x_i)$ or equivalently its logarithm, the log-likelihood $L(f, x_1, \ldots, x_n)$:

$$\hat{f} := \operatorname{argmax}_{f \in \mathcal{F}} L(f, x_1, \ldots, x_n) := \operatorname{argmax}_{f \in \mathcal{F}} \sum_{i=1}^{n} \log(f(x_i)).$$

Without further restrictions on the class $\mathcal{F}$, the log-likelihood is unbounded, and hence, no maximum likelihood estimate exists. One possibility is to

restrict $\mathcal{F}$ to a set of histograms. Consider a partition $\mathcal{I} := \{I_1, \ldots, I_D\}$ of a compact interval $K \subset \mathbb{R}$ into D intervals $I_1, \ldots, I_D$, such that $I_i \cap I_j = \emptyset$ for $i \neq j$ and $\bigcup I_i = K$. Now consider the set $\mathcal{F}_\mathcal{I}$ of all histograms that are piecewise constant on $\mathcal{I}$ and zero outside $\mathcal{I}$:

$$\mathcal{F}_\mathcal{I} := \left\{ f \,\middle|\, f = \sum_{j=1}^{D} h_j \mathbf{1}_{I_j}, h_j \geq 0, j = 1, \ldots, D \text{ and } \sum_{j=1}^{D} h_j |I_j| = 1 \right\},$$

where $\mathbf{1}_A$ denotes the indicator function of a set A and $|I|$ the length of the interval I. If K contains $[x_{(1)}, x_{(n)}]$, the *Maximum Likelihood Histogram* (ML histogram) is defined as the maximizer of the log-likelihood in $\mathcal{F}_\mathcal{I}$ and is given by

$$\hat{f}_\mathcal{I} := \operatorname{argmax}_{f \in \mathcal{F}_\mathcal{I}} L(f, x_1, \ldots, x_n) = \frac{1}{n} \sum_{j=1}^{D} \frac{N_j}{|I_j|} \mathbf{1}_{I_j}, \tag{1}$$

with $N_j = \sum_{i=1}^{n} \mathbf{1}_{I_j}(x_i)$. Its log-likelihood is

$$L(\hat{f}_\mathcal{I}, x_1, \ldots, x_n) = \sum_{j=1}^{D} N_j \log \frac{N_j}{n|I_j|}. \tag{2}$$

In the following, we consider partitions $\mathcal{I} := \mathcal{I}_D := (I_1, \ldots, I_D)$ of the interval $I := [x_{(1)}, x_{(n)}]$, consisting of D intervals of the form

$$I_j := \begin{cases} [t_0, t_1] & j = 1 \\ (t_{j-1}, t_j] & j = 2, \ldots, D \end{cases},$$

with breakpoints $x_{(1)} =: t_0 < t_1 < \cdots < t_D := x_{(n)}$. A histogram is called *regular* if all intervals have the same length and *irregular* otherwise. The intervals are also referred to as *bins*.

We will only consider ML histograms in this work, and use the term "histogram" synonymously with "ML histogram" unless explicitly stated otherwise. We focus on finding a data-driven construction of a histogram with

good risk behavior. Given a distance measure d between densities, the risk is defined as the expected distance between the true and the estimated density:

$$R_n(f, \hat{f}_{\mathcal{I}}, d) := E_f[d(f, \hat{f}_{\mathcal{I}}(X_1, \ldots, X_n))].$$

We consider the risks w.r.t. (normalized) squared Hellinger distance

$$d_H(f, g) = \frac{1}{2} \int (\sqrt{f(t)} - \sqrt{g(t)})^2 dt, \tag{3}$$

and w.r.t. powers of the L_p-norms ($p = 1, 2$) defined by

$$d_p := \|f - g\|_p^p = \int |f(t) - g(t)|^p dt. \tag{4}$$

For a detailed discussion on the choice of loss functions in histogram density estimation, see Birgé and Rozenholc (2006, sec. 2.2) and the references given there.

Given the sample, the histogram $\hat{f}_{\mathcal{I}}$ depends only on the partition $\mathcal{I} = (I_1, \ldots, I_D)$ as the values on the intervals of the partition are given by (1). In order to achieve good performance in terms of risk, the crucial point is thus choosing the partition. Since partitions with too many bins will result in a large likelihood without yielding a good estimate of f, a naïve comparison of the likelihoods of histograms with different numbers of bins is misleading. But also without any further restrictions on the allowed partitions the likelihood can be made arbitrarily large for a fixed number of bins.

Many approaches exist for the special case of regular histograms where I is divided into D equal sized bins; the problem is then reduced to the choice of D, cf. Birgé and Rozenholc (2006), Davies et al. (2009) and the references given there. Using irregular partitions can reduce bias and therefore can improve performance for spatially inhomogenous densities, but the increased

difficulty of choosing a good partition may lead to an increase in risk for more well-behaved densities. The idea of constructing both a regular and an irregular histogram and then choosing between the two is briefly discussed in Birgé and Rozenholc (2006). To our knowledge, this approach has not yet been put into practice.

Our recommendation is to construct a regular histogram that maximizes the penalized log-likelihood

$$L(\hat{f}_{\mathcal{I}}, x_1, \ldots, x_n) - (D-1) - \log^{2.5} D \tag{5}$$

among all regular partitions of $[x_{(1)}, x_{(n)}]$ with $D = 1, \ldots, \lfloor n/\log n \rfloor$ bins (where $\lfloor x \rfloor$ denotes the largest integer not larger than x) and an irregular histogram that maximizes the penalized log-likelihood

$$L(\hat{f}_{\mathcal{I}}, x_1, \ldots, x_n) - \log \binom{n-1}{D-1} - (D-1) - \log^{2.5} D \tag{6}$$

among a set of partitions of $[x_{(1)}, x_{(n)}]$ with breakpoints equal to the sample points, where again D is the number of bins in a given partition. The final estimate is then the one with the larger penalized log-likelihood. The penalty in (5) for the regular case was proposed in Birgé and Rozenholc (2006), while the motivation for (6) is developed later in this paper, where we consider different penalty forms for the irregular case. Note that the difference between the penalties is the term $\log \binom{n-1}{D-1}$ which is needed because in the irregular case, the best partition with D bins has to be chosen from a set of $\binom{n-1}{D-1}$ partitions, while there is only one partition with D bins in the regular case. The necessity of taking into account in the penalty not only the number of parameters in a model but also the number of candidate models with same number of parameters is one of the key points in Barron

et al. (1999). Specific penalty forms for histogram estimators were derived in Castellan (1999) and for more general situations in Massart (2007). Note that both penalties (and hence the penalized log-likelihoods) coincide for $D = 1$. The penalties are both designed to achieve a good control on the Hellinger risk in term of Oracle Inequality as derived for example in (Massart, 2007, Th.??? eq(???)).**Thoralf, I don't have the Massart book with me, could you extract an Oracle Inequality from a theorem connected to histogram**

Several methods for choosing a good irregular histogram have been developed previously. Kogure (1987) gives asymptotic results for the optimal choice of bins. His approach is based on using blocks of equisized bins, and he explores the dependence on tuning parameters via simulations (Kogure, 1986). It does not result in a fully automatic procedure. Kanazawa (1988) proposes to control the Hellinger distance between the unknown true density and the estimated histogram and introduces a dynamic programming algorithm to find the best partition with a given number of bins. Kanazawa (1992) derives the asymptotically optimal choice of the number of bins which depends on derivatives of the unknown density, making the procedure inapplicable in practice. Celisse and Robin (2008) give explicit formulas for L_2 leave-p-out cross-validation for regular and irregular histograms. They only briefly comment on the case of irregular histograms and only show simulations with ad-hoc choices of the set of partitions. In our simulations, we use their explicit formula to compare risk behavior of cross-validation and our penalized likelihood approach when both are used to choose an irregular histogram from the same set of partitions. The multiresolution histogram by

Engel (1997) is based on a tree of dyadic partitions. Its performance crucially depends on the finest resolution level, for which no universally usable recommendation is given. Some other tree-based procedures have been suggested for the multivariate case. They can be used for the univariate case, but they either perform a complete search over a restricted set of partitions (Blanchard et al., 2007; Klemelä, 2009) or a greedy search over a full set of partitions (Klemelä, 2007) to deal with computational problems that do not occur in the univariate case. Conditions for consistency of histogram estimates with data-driven and possibly irregular partitions are derived in Chen and Zhao (1987); Lugosi and Nobel (1996) and Zhao et al. (1988). Devroye and Lugosi (2004) give a construction of histograms where bin widths are allowed to vary according to a pre-specified function.

Hartigan (1996) considers regular and irregular histogram construction from a Bayesian point of view. However, we are not aware of any fully tuned automatic Bayesian procedure for irregular histogram construction. Rissanen et al. (1992) give a construction based on the Minimum Description Length (MDL) paradigm, which leads to a penalized likelihood estimator. A choice of several discretization parameters is needed, and the recommendation given by the authors is to perform an exhaustive search over all possible combinations of values, which is computationally expensive. A more recent MDL-based proposal by Kontkanen and Myllymäki (2007) involves a discretization which results in the estimate not being a proper density. Catoni (2002) suggests a multi-stage procedure based on coding ideas that computes a density estimate by aggregating histograms.

The taut string procedure introduced by Davies and Kovac (2004) can

also be used to generate an irregular histogram as described in Davies et al. (2009). Regularization is achieved not by controlling the number of bins but by constructing an estimate that has a minimum number of modes subject to a constraint on the distance between the empirical and the estimated distribution function. The main idea is to construct a piecewise linear spline of minimal length (the taut string) in a tube around the empirical cdf and then take its derivative, which is piecewise constant. With some modifications this gives a histogram that fulfills definition (1). The main tuning parameter is the tube width, and an automatic choice is suggested by the authors. Although not designed to minimize risk, the procedure has performed well w.r.t. classical loss functions (Davies et al., 2009), and therefore is included in our simulations.

For our construction of irregular histograms, we will focus on penalized likelihood maximization techniques. For a good data-driven histogram one needs an appropriate penalization to provide an automatic choice of D as well as of the partition $\mathcal{I} = (I_1, \ldots, I_D)$. Since Akaike's Information Criterion (AIC) introduced by Akaike (1973), penalized likelihood has been used with many different penalty terms. AIC aims at ensuring a good risk behavior of the resulting estimate. Another widely used criterion is the Bayesian Information Criterion (BIC) introduced by Schwarz (1978). It is constructed in such a way as to consistently estimate the smallest true model order, which in histogram density estimation would lead to very large models unless the true density is piecewise constant. In practice, criteria like AIC and BIC are routinely applied in many different statistical models, often without reference to their different conceptual backgrounds and without appropriate

modifications for the model under consideration. In their original forms, both AIC and BIC do not account for multiple partitions with the same number of bins. See Chapter 7.3 of Massart (2007) for a critique of the use of AIC in histogram density estimation. Since both are widely used, we include them in our comparisons. Our penalties are motivated by recent model selection works by Barron et al. (1999), Castellan (1999, 2000) and Massart (2007). The regular histogram construction proposed in Birgé and Rozenholc (2006) with which we combine our irregular histogram is based on the same ideas.

Our paper is structured as follows: In Section 2, we review the problem of constructing an irregular histogram using penalized likelihood. Section 3 gives a description of the practical implementation including calibration of the penalty. Section 4 gives the results of a simulation study and conclusions.

2. Penalized likelihood construction of irregular histograms

Constructing an irregular histogram by penalized likelihood means maximizing

$$L(\hat{f}_{\mathcal{I}}, x_1, \ldots, x_n) - \mathrm{pen}_n(\mathcal{I}), \tag{7}$$

w.r.t. partitions $\mathcal{I} = (I_1, \ldots, I_{|\mathcal{I}|})$ of $[x_{(1)}, x_{(n)}]$, where $\mathrm{pen}_n(\mathcal{I})$ is a penalty term depending only on the partition $\mathcal{I}$ and possibly on the sample (data-driven). We will introduce a new choice here motivated by work of Barron et al. (1999), Castellan (1999, 2000) and Massart (2007).

Optimizing w.r.t. the partition $\mathcal{I}$ with $|\mathcal{I}|$ fixed in (7) leaves us with a continuous optimization problem. Without further restrictions, for $|\mathcal{I}| \geq 2$ the likelihood is unbounded. One possibility is to restrict to all partitions

that are built with endpoints on the observations; the optimization problem (7) can then be solved using a dynamic programming algorithm first used for histogram construction by Kanazawa (1988). More details are given in Section 3.

With $D = |\mathcal{I}|$, we propose the following families of penalties parametrized by two constants c and α:

$$\text{pen}_n^A(\mathcal{I}) = c \log \binom{n-1}{D-1} + \alpha(D-1) + \varepsilon_{c,\alpha}^{(1)}(D), \quad (8)$$

$$\text{pen}_n^B(\mathcal{I}) = c \log \binom{n-1}{D-1} + \alpha(D-1) + \varepsilon^{(2)}(D) \quad (9)$$

and

$$\text{pen}_n^R(\mathcal{I}) = c \log \binom{n-1}{D-1} + \frac{\alpha}{n} \sum_{j=1}^{D} \frac{N_j}{|I_j|} + \varepsilon^{(2)}(D), \quad (10)$$

where

$$\varepsilon_{c,\alpha}^{(1)}(D) = c\,k \log D + 2\sqrt{c\alpha(D-1)(\log \binom{n-1}{D-1} + k \log D)} \quad (11)$$

and

$$\varepsilon^{(2)}(D) = \log^{2.5} D. \quad (12)$$

The precise choices for c and α obtained by simulations are described in Section 3. Note that, while the penalties given in (8) and (9) depend only on the number of bins of the partition, the penalty in formula (10) is a *random penalty* in the sense that it also depends on the data.

We now give arguments to explain the origins of these penalties. The penalty defined by (8) is derived from Theorem 3.2 in Castellan (1999), which is also stated as Theorem 7.9 in Massart (2007, p. 232) and from eq. (7.32)

in Theorem 7.7 in Massart (2007, p.219). From the penalty form in Theorem 7.9 in Massart (2007) we derive $\varepsilon^{(1)}$:

$$\text{pen}_n(\mathcal{I}) = c_1(\sqrt{D-1} + \sqrt{c_2 x_{\mathcal{I}}})^2, \tag{13}$$

where the weights $x_{\mathcal{I}}$ are chosen such that

$$\sum_D \sum_{|\mathcal{I}|=D} e^{-x_{\mathcal{I}}} \leq \Sigma \tag{14}$$

for an absolute constant Σ. Because the endpoints of our partitions are fixed, there are $\binom{n-1}{D-1}$ different partitions with cardinality D, and we assign equal weights to every partition $\mathcal{I}$ with $|\mathcal{I}| = D$:

$$x_{\mathcal{I}} = \log \binom{n-1}{D-1} + k \log D.$$

Choosing $k > 1$ ensures that the sum in (14) is converging and that Σ is finite. Substitution into (13) gives

$$\begin{aligned} \text{pen}_n(\mathcal{I}) &= c_1 \left(D - 1 + c_2 \left(\log \binom{n-1}{D-1} + k \log D \right) \right. \\ &\quad \left. + 2\sqrt{c_2(D-1)\left(\log \binom{n-1}{D-1} + k \log D \right)} \right). \end{aligned} \tag{15}$$

Let us emphasize that Theorem 7.9 in Massart (2007, p. 232) requires $c_1 > 1/2$ and $c_2 = 2(1 + 1/c_1)$. Coming back to our notations, with $\alpha = c_1$, $c = c_1 c_2$ we obtain Equation (11).

We now use Theorem 7.7 in Massart (2007, p. 219) to justify the random penalty in (10). The orthonormal basis considered in this theorem for a given partition $\mathcal{I}$ consists of all $\mathbf{1}_I/\sqrt{|I|}$ for all I in $\mathcal{I}$. The least squares contrast used in this theorem in our framework is $-n^{-2}\sum_{I\in\mathcal{I}} N_I^2/|I|$. To

link the minimization of the least squares contrast and the maximization of the log-likelihood, we consider the following approximation:

$$L(\hat{f}_{\mathcal{I}}, x_1, \ldots, x_n) = \sum_{j=1}^{D} N_j \log\left(\frac{N_j}{n|I_j|}\right) \approx \sum_{j=1}^{D} N_j \left(\frac{N_j}{n|I_j|} - 1\right) = \frac{1}{n}\sum_{j=1}^{D} \frac{N_j^2}{|I_j|} - n.$$

From the penalty form (7.32) and the use of $M = 1$ and $\varepsilon = 0$ in Theorem 7.7 in Massart (2007, p. 219), following the same derivation for $\varepsilon^{(1)}$, we find the penalty in (15) with $c_1 = 1$ and $c_2 = 2$.

Using the least squares approximation, we can use the random penalty (7.33) in Theorem 7.7 in Massart (2007). Let us emphasize that $\widehat{V}_m$ defined by Massart is in our framework $\sum_{I\in\mathcal{I}} N_I/n|I|$ with $m = \mathcal{I}$. To derive $\varepsilon^{(2)}$ in (10) we start from the penalty defined in (7.33) in Massart (2007):

$$\mathrm{pen}_n(\mathcal{I}) = (1+\varepsilon)^5 \left(\sqrt{\widehat{V}_{\mathcal{I}}} + \sqrt{2ML_{\mathcal{I}}D}\right)^2.$$

Following the same derivations as for the penalty (13), setting $M = 1$, $\varepsilon = 0$ and $L_{\mathcal{I}} = D^{-1}(\log\binom{n-1}{D-1} + k\log D)$ we obtain:

$$\begin{aligned}\mathrm{pen}_n(\mathcal{I}) &= \widehat{V}_{\mathcal{I}} + 2\log\binom{n-1}{D-1} + 2k\log D \\ &\quad +2\sqrt{2\widehat{V}_{\mathcal{I}}\left(\log\binom{n-1}{D-1} + k\log D\right)}.\end{aligned}$$

Let us emphasize that, because of terms of the form $\varphi(D)\widehat{V}_{\mathcal{I}}$, the expression in the square root above prevents the use of dynamic programming to compute the maximum of the penalized log-likelihood defined in (7). To avoid this problem we propose, following penalty forms proposed in Birgé and Rozenholc (2006) and Comte and Rozenholc (2004), to replace the remainder expression $2k\log D + 2\sqrt{2\widehat{V}_{\mathcal{I}}\left(\log\binom{n-1}{D-1} + k\log D\right)}$ by a power of

$\log D$. We have tried several values of the power and found that formula (12) leads to a good choice. Finally, we also replaced $\varepsilon_{c,\alpha}^{(1)}$ in formula (8) by $\varepsilon^{(2)}$, leading to the penalty given in (9).

3. Practical Implementation

We will describe briefly the implementation of our method. For a more detailed description, see Rozenholc et al. (2009). To calibrate the penalties, histograms with the endpoints of the partitions placed on the observations and with different choices of the constants α and c in the penalties given in (8), (9) and (10) were evaluated by means of simulations. We used the same densities for calibration as in the simulations described in Section 4 but different samples and a smaller number of replications. The loss functions $d = d_H, d_1, d_2$ were evaluated by numerical integration using a trapeze rule. We focused on the Hellinger risk to obtain good choices of the penalties, but the behavior w.r.t. L_1 loss is very similar. For minimizing L_2 risk, other choices may be preferable. Since no single penalty is best in all cases, we describe in the following what we consider to be a good compromise.

In formula (8) we tried different combinations of $\alpha \in \{0.5, 1\}$ and c between 1 and 4, some of which were motivated by Theorems 7.7 (eq. 7.32) and 7.9 in Massart (2007). We always set $k = 2$. From these experiments, the most satisfactory choice is $c = 1$ and $\alpha = 0.5$. We also ran experiments replacing $\varepsilon_{c,\alpha}^{(1)}$ by $\varepsilon^{(2)}$, leading to the penalty given in (9). In this case, the most satisfactory choice is $c = 1$ and $\alpha = 1$, and this choice is even better than $\varepsilon_{2,1}^{(1)}$. Note that the resulting penalty, given in (6), exactly corresponds to the penalty in (5) proposed in Birgé and Rozenholc (2006) for the regu-

lar case, except for the additional term $\log\binom{n-1}{D-1}$ that is needed to account for multiple partitions with the same number of bins. This term vanishes for $D = 1$, making the maxima of penalized likelihoods directly comparable. Because (8) and (9) are very similar, we only use this version in our simulations in Section 4.

For the random penalty in formula (10) we tried all combinations of $c \in \{0.5, 1, 2\}$ and $\alpha \in \{0.5, 1\}$. Let us emphasize that $c = 2$ and $\alpha = 1$ correspond to formula (7.33) in Massart (2007) up to our choice of $\varepsilon^{(2)}$ defined in (12). From our point of view, the most satisfactory choice is $c = 1$ and $\alpha = 0.5$. In order to make the maximum of the log-likelihood penalized by (10) comparable to the maximum of (5), we add the constant $\alpha(x_{(n)} - x_{(1)})$, which does not change the maximizer but makes the penalized log-likelihoods coincide for $D = 1$.

We now briefly describe the algorithm for constructing the irregular histogram. We consider partitions $\mathcal{I}$ built with endpoints on the observations:

$$\mathcal{I} = ([x_{(1)}, x_{(k_1)}], (x_{(k_1)}, x_{(k_2)}], (x_{(k_2)}, x_{(k_3)}], \ldots, (x_{(k_{D-2})}, x_{(k_{D-1})}], (x_{(k_{D-1})}, x_{(n)}]),$$

where $1 < k_1 < \ldots < k_{D-1} < n$. We start from a "finest" partition $\mathcal{I}_{\max}$ defined by $D_{\max} < n$ and the choice $1 < k_1 < \ldots < k_{D_{\max}-1} < n$. Our aim is to construct a sub-partition $\mathcal{I}$ of $\mathcal{I}_{\max}$ that maximizes (7). For all penalties (8)-(10) this can be achieved by a dynamic programming algorithm (Kanazawa, 1988; Comte and Rozenholc, 2004). The total complexity of the algorithm is of order $D_{\max}^3$. We reduce this to the order n by first using a greedy algorithm to construct a partition with $\lfloor \max\{n^{1/3}, 100\} \rfloor$ bins if this number is smaller than n. Starting with a partition with one bin, in each step we recursively add to the current partition an endpoint (at an observation)

in an existing bin in order to achieve the best maximization (in one step) of the likelihood. The resulting partition is used as the finest partition for the dynamic programming algorithm.

Let us remark that the theoretical results by Castellan (1999, 2000) and Massart (2007, ch. 7), are derived for the case of a finest regular grid (depending on n but not on the sample) with bin sizes not smaller than a constant times $\log^2(n)/n$. However, we found that, in practice, we can improve performance drastically for some densities by using a data-dependent finest grid imposing no restrictions on the smallest bins without losing much at other densities. More comments on this are given in Section 4.

4. Simulation Study and Conclusions

The performance of our proposals given in Section 3 is compared to the performance of other available methods for 12 of the 28 test-bed densities introduced by Berlinet and Devroye (1994) which include standard distributions like the uniform and the normal as well as more special ones. We denote these by $f_1, \ldots, f_{12}$. We also add 4 histogram densities $f_{13}, \ldots, f_{16}$ which are defined in Rozenholc et al. (2009). Some of the densities do not fulfill the conditions in Theorem 3.2 in Castellan (1999) for example if they are not bounded away from zero. They are included in order to explore the behavior of the procedure in cases not covered by theory. All densities are implemented in the `R`-package `benchden` (Mildenberger et al., 2009b) and are depicted in Figure 1.

The sample sizes are 50,100,500,1000,5000 and 10000. We use 500 replications for each scenario. The methods compared in the simulations are:

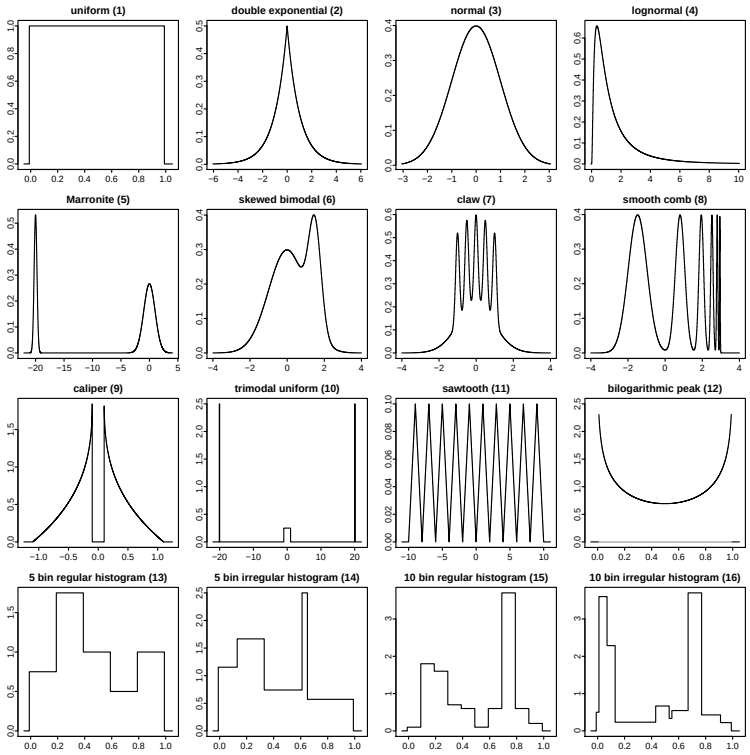

Figure 1: The densities used in the simulation study.

- **B** Penalized maximum likelihood using penalty (9) with $c = 1$ and $\alpha = 1$.

  R Penalized maximum likelihood using random penalty (10) with $c = 1$ and $\alpha = 0.5$.

  CV Leave-one-out cross-validation using formula (11) given in Celisse and Robin (2008). We also tried formula (12) of Celisse and Robin (2008) for different values of p without finding a big difference.

Maximization is performed over a data-driven finest grid as described in Section 3 without restrictions on the minimum bin width.

- Methods $\mathbf{B}, \mathbf{R}, \mathbf{CV}$ using the same data-driven grid but with the additional constraint that the minimum bin length allowed is $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$. These are denoted by $\mathbf{B}'$, $\mathbf{R}'$ and $\mathbf{CV}'$.

- Methods $\mathbf{B}, \mathbf{R}, \mathbf{CV}$ but using a full optimization over a finest regular partition with bin width $(x_{(n)} - x_{(1)})\lfloor n/\log^{1.5}(n)\rfloor^{-1}$. This is the grid considered in Castellan (1999), except that we slightly relax $\log^2(n)$ to $\log^{1.5}(n)$. These are denoted $\mathbf{B}''$, $\mathbf{R}''$ and $\mathbf{CV}''$.

- Penalized maximum likelihood using Information Criteria:

  AIC Akaike's Information Criterion (Akaike, 1973). The penalty is $\mathrm{pen}_n^{\mathrm{AIC}}(D) = (D-1)$.

  BIC Bayesian Information Criterion (Schwarz, 1978). The penalty is $\mathrm{pen}_n^{\mathrm{BIC}}(D) = 0.5\log(n)(D-1)$.

- The taut-string method **TS** introduced by Davies and Kovac (2004). We use the function `pmden()` implemented in the R-package `ftnonpar` (Davies and Kovac, 2008) with the default values except that we set `localsq=FALSE` as local squeezing of the tube does not give a ML histogram. The histogram is then constructed using the knots of the string as the boundaries of the bins. This coincides with the derivative of the string except on intervals where the string switches from the upper to the lower boundary of the tube or vice versa.

- Regular histogram construction **BR** due to Birgé and Rozenholc (2006). The penalty is $\mathrm{pen}_n^{\mathrm{BR}}(D) = (D-1)+\log(D)^{2.5}$, where the log-likelihood is maximized over all regular partitions with $1, \ldots, \lfloor n/\log n\rfloor$ bins.

- Combined approach: A regular histogram and an irregular histogram are constructed. After adding a constant such that all penalties are 0 for a histogram with 1 bin, the penalized log-likelihoods are compared. Such a procedure mixing several collections of models was proposed in Barron et al. (1999) but, to our knowledge, it has never been applied from a practical point of view. The aim is to achieve the best trade-off between bias and variance. Indeed, very large collections like those consisting of irregular histograms may lead to a small bias but may increase the stochastic part of the risk. The necessary price are terms of order $\log \binom{n-1}{D-1}$ in the penalty which lead to larger upper bounds on the risk in the oracle inequality. It is then of interest to compare more than one strategy: one giving good results when the density is well-behaved (using the regular histogram collection) and one for less well-behaved (e.g. less smooth) densities (using the irregular histogram collection). Because the density is unknown and so is its regularity, it is proposed to directly compare the penalized likelihoods in each collection. The histogram which has the larger penalized likelihood is chosen as the final estimate:

  B* Chooses between a regular histogram using **BR** and an irregular histogram using **B**

  R* Chooses between a regular histogram using **BR** and an irregular histogram using **R**.

Except for **TS**, all methods have been implemented in the `R`-package `histogram` (Mildenberger et al., 2009a). The program code to reproduce all simulation

results is available from the corresponding author's website. For the discussion of the results, we focus on squared Hellinger risk. Table 1 shows the binary logarithms of relative squared Hellinger risks w.r.t. the best method for any given n and density: $\log_2(\widehat{R}_n^{\text{method}}/\widehat{R}_n^{\text{best}})$ for all 15 methods in the simulation study. Thus, a value of 0 means that the method was best in this particular setting and a value of 1 means that the risk of the method is twice as large as the risk of the best method. We omitted $n = 5000$ for space reasons; complete tables giving the absolute risk values, tables for other risks and a more detailed discussion of the results can be found in Rozenholc et al. (2009). Generally, the L_1 and Hellinger risks behave similarly, while the results for L_2 are much less conclusive and no clear recommendation for using one of the methods under consideration can be given when minimizing the L_2 risk is the main aim.

As was to be expected, the table shows no clear overall best method for all scenarios. The relative risk w.r.t. the best method is always smaller than 2 for our proposal $\mathbf{B}^*$. In this sense, it can be seen as the best method. Let us remark that other methods could be better in particular situations. In many cases, **TS** or one of our proposals **B** and **R** is either the best or the binary logarithm of relative risk w.r.t. the best method is close to zero. These three methods are – except for $\mathbf{B}^*$ – also the only ones in the simulation study for which this quantity is always strictly smaller than $\log_2 3 \approx 1.58$, meaning that the empirical risk is never greater than three times the risk achieved by the best method. The random penalty **R** seems to be slightly better than **B** in many cases, the most notable exception being the trimodal uniform density f_{10} for $n = 50$.

	n	**B**	**R**	**CV**	**B′**	**R′**	**CV′**	**B″**	**R″**	**CV″**	**AIC**	**BIC**	**TS**	**BR**	**B***	**R***
f_1	50	0.13	0.05	1.53	0.03	0.03	0.76	0	0	0.44	2.07	1.27	0.07	0.12	0.18	0.13
	100	0.08	0.02	2.31	0.02	0.02	1.25	0	0.01	0.72	2.91	1.56	0.1	0.11	0.17	0.12
	500	0.06	0.03	4.2	0.01	0.02	2.56	0	0.01	1.69	4.64	2.04	0.19	0.13	0.17	0.14
	1000	0.04	0.02	4.64	0.01	0.01	3.05	0	0.01	2.3	4.98	2.1	0.12	0.11	0.14	0.12
	10000	0	0	5.39	0	0	3.98	0	0	4.71	5.59	1.85	0.2	0.18	0.19	0.18
f_2	50	0.42	0.34	0.5	0.38	0.34	0.09	0.52	0.47	0	0.91	0.53	0.13	0.14	0.17	0.16
	100	0.57	0.51	0.81	0.56	0.51	0.05	0.56	0.54	0	1.34	0.61	0.02	0.2	0.24	0.25
	500	0.71	0.67	1.79	0.7	0.66	0.04	0.69	0.67	0.07	2.2	0.75	0	0.2	0.2	0.2
	1000	0.74	0.7	1.94	0.72	0.7	0.08	0.72	0.7	0.15	2.25	0.73	0	0.18	0.18	0.18
	10000	0.69	0.71	1.59	0.69	0.71	0.21	0.67	0.68	0.68	1.7	0.56	0	0.07	0.07	0.07
f_3	50	0.51	0.42	0.68	0.49	0.43	0.14	0.51	0.48	0	1.15	0.64	0.51	0.23	0.24	0.23
	100	0.47	0.4	1.07	0.45	0.39	0.17	0.48	0.43	0	1.62	0.72	0.25	0.15	0.16	0.17
	500	0.65	0.6	2.08	0.64	0.6	0.36	0.65	0.59	0.1	2.49	0.81	0.13	0	0	0
	1000	0.73	0.69	2.28	0.73	0.69	0.54	0.73	0.66	0.25	2.59	0.84	0.19	0	0	0
	10000	1.03	0.99	2.19	1.02	0.99	0.94	1	0.96	1.34	2.32	0.93	0.57	0	0	0
f_4	50	0.56	0.5	0.68	0.74	0.72	0.45	0.68	0.65	0.41	1.12	0.67	0	0.55	0.56	0.52
	100	0.63	0.6	1.08	0.72	0.72	0.38	0.7	0.69	0.38	1.62	0.79	0	0.61	0.62	0.59
	500	0.56	0.61	1.79	0.71	0.72	0.36	0.66	0.69	0.41	2.2	0.62	0	0.67	0.55	0.62
	1000	0.31	0.42	1.74	0.72	0.74	0.4	0.7	0.73	0.46	2.05	0.39	0	0.75	0.34	0.52
	10000	0.55	0.77	1.62	1.48	1.56	1.27	1.27	1.37	1.25	1.73	0.45	0	2.05	0.55	0.77
f_5	50	0.25	0.19	0.5	2.18	2.18	2.18	1.6	1.59	1.57	0.64	0.33	0	1.12	0.25	0.19
	100	0.47	0.44	0.76	2.71	2.71	2.73	1.81	1.82	1.77	1.07	0.48	0	0.41	0.47	0.44
	500	0.79	0.77	1.73	1.69	1.69	1.92	1.89	1.9	1.68	2.05	0.81	0	1.18	0.79	0.77
	1000	0.89	0.91	1.92	1.56	1.56	1.78	1.29	1.3	0.98	2.16	0.85	0	0.95	0.89	0.91
	10000	0.98	1	1.63	0.98	1	0.47	0.96	1	0.22	1.71	0.82	0	0.68	0.98	1
f_6	50	0.4	0.31	0.75	0.38	0.32	0.24	0.33	0.3	0	1.22	0.66	0.51	0.28	0.3	0.27
	100	0.36	0.29	1.14	0.35	0.3	0.31	0.45	0.39	0	1.7	0.8	0.44	0.31	0.31	0.29
	500	0.5	0.45	1.98	0.5	0.45	0.31	0.48	0.42	0	2.38	0.71	0.1	0.09	0.09	0.1
	1000	0.55	0.49	2.11	0.54	0.49	0.42	0.52	0.47	0.11	2.42	0.67	0.06	0	0	0.01
	10000	0.92	0.87	2.1	0.91	0.87	0.85	0.91	0.86	1.26	2.23	0.83	0	0.11	0.11	0.11
f_7	50	0.28	0.16	0.48	0.25	0.17	0.04	0.39	0.35	0	0.86	0.45	0.39	0.19	0.2	0.15
	100	0.05	0.01	0.66	0.02	0	0.08	0.2	0.18	0.05	1.15	0.53	0.06	0.14	0.12	0.08
	500	0.68	0.64	1.44	0.66	0.63	0.14	0.69	0.66	0.15	1.84	0.68	0	0.27	0.28	0.3
	1000	0.89	0.87	1.8	0.88	0.86	0.44	1.15	1.14	0.29	2.1	0.8	0	0.45	0.45	0.45
	10000	1.25	1.24	1.85	1.25	1.24	0.76	1.26	1.26	1.11	1.94	1.11	0	0.77	0.77	0.77
f_8	50	0.3	0.24	0.08	0.29	0.26	0.05	0.34	0.32	0	0.4	0.21	0.22	0.14	0.16	0.14
	100	0.37	0.28	0.27	0.37	0.3	0	0.63	0.55	0.09	0.67	0.27	0.17	0.09	0.09	0.09
	500	0.44	0.36	0.9	0.51	0.46	0.24	0.66	0.59	0	1.25	0.35	0.03	0.1	0.1	0.1
	1000	0.47	0.39	1.05	0.54	0.49	0.27	0.64	0.58	0.09	1.33	0.37	0	0.19	0.19	0.19
	10000	0.99	0.93	1.17	1.01	0.96	0.61	0.99	0.93	0.66	1.25	0.82	0	0.7	0.7	0.7
f_9	50	0.44	0.42	0	0.45	0.45	0.16	0.44	0.44	0.36	0.34	0.03	0.26	0.36	0.38	0.37
	100	0.46	0.28	0.72	0.49	0.28	0	1.37	1.34	1.09	1.13	0.47	0.17	0.6	0.51	0.37
	500	0.55	0.5	1.81	0.52	0.48	0.36	1.64	1.62	1.36	2.17	0.68	0	0.85	0.63	0.55
	1000	0.61	0.56	2.02	0.59	0.55	0.5	1.65	1.62	1.37	2.3	0.72	0	1.08	0.66	0.61
	10000	0.73	0.69	1.87	0.72	0.69	0.74	1.61	1.59	1.69	1.98	0.65	0	1.31	0.77	0.74
f_{10}	50	0	1.31	0.55	3.15	3.15	3.14	3.12	3.12	3.12	0.68	0.31	0.15	2.68	0	2.68
	100	0.06	0	1.02	4.2	4.2	4.2	4.02	4.02	4.02	1.32	0.53	0.32	3.29	0.06	0
	500	0.02	0	2.4	5.82	5.82	5.82	5.64	5.64	5.64	2.74	0.79	0.51	5.08	0.02	0
	1000	0.01	0	2.8	6.8	6.8	6.8	6.34	6.34	6.34	3.07	0.82	0.56	5.63	0.01	0
	10000	0	0	3.48	10.1	10.1	10.1	6.46	6.46	6.46	3.59	0.65	0.81	0.17	0	0
f_{11}	50	0.06	0.02	0.54	0.01	0.01	0.22	0	0	0.13	0.81	0.51	0.19	0.03	0.08	0.04
	100	0.02	0.01	0.39	0	0	0.24	0	0	0.13	0.74	0.43	0.25	0.03	0.05	0.03
	500	0.86	0.61	0.8	0.89	0.7	0.27	1.06	1.06	0.15	1.14	0.48	0.11	0	0	0
	1000	0.63	0.58	0.98	0.62	0.59	0.4	0.86	0.69	0.08	1.24	0.63	0.09	0	0	0
	10000	1.06	0.99	0.93	1.05	0.98	0.57	1.09	1.01	0.37	0.97	0.91	0.17	0	0	0
f_{12}	50	0.21	0.11	0.88	0.01	0.02	0.25	0	0.01	0.14	1.36	0.79	0.03	0.15	0.27	0.18
	100	0.23	0.17	1.17	0.13	0.14	0.27	0.12	0.13	0	1.75	0.73	0.15	0.16	0.24	0.18
	500	0.37	0.33	2.41	0.3	0.3	0.88	0.29	0.28	0.28	2.83	0.9	0	0.29	0.37	0.34
	1000	0.52	0.49	2.63	0.48	0.47	1.09	0.52	0.51	0.5	2.95	0.89	0	0.36	0.45	0.42
	10000	0.68	0.66	2.48	0.67	0.66	1.18	0.67	0.65	1.61	2.62	0.74	0	0.57	0.6	0.59
f_{13}	50	0.1	0.07	0.69	0.04	0.05	0.02	0.01	0.02	0	1.16	0.65	0.06	0.06	0.09	0.06
	100	0.43	0.4	1.21	0.39	0.38	0.25	0.38	0.38	0.04	1.77	0.87	0.4	0	0.01	0.01
	500	1.41	1.34	3.43	1.39	1.33	1.86	1.53	1.5	1.41	3.85	1.88	1.1	0	0	0
	1000	1.06	0.98	3.98	1.02	0.98	2.44	0.83	0.78	1.7	4.29	1.97	1.11	0	0	0
	10000	0.92	0.92	4.83	0.9	0.91	3.57	1.39	1.39	4.1	5	1.93	1.25	0	0	0
f_{14}	50	0.18	0.13	0.65	0.12	0.11	0.14	0.12	0.13	0	1.1	0.59	0.19	0.02	0.08	0.04
	100	0.26	0.23	1	0.24	0.23	0.3	0.36	0.33	0.19	1.54	0.65	0.29	0	0.03	0.02
	500	0.08	0.03	2.41	0.02	0	0.85	0.58	0.56	0.67	2.83	0.83	0.09	0.88	0.35	0.28
	1000	0.14	0.11	2.98	0.12	0.11	1.45	0.83	0.82	1.25	3.3	0.98	0	0.94	0.21	0.15
	10000	0.01	0.01	3.96	0	0	2.67	0.92	0.93	3.31	4.13	1.04	0.57	2.64	0.01	0.01
f_{15}	50	0.1	0	0.06	0.35	0.29	0.06	0.46	0.42	0.06	0.43	0.09	0.11	0.03	0.04	0.01
	100	0.27	0.18	0.65	0.24	0.21	0	0.69	0.65	0.35	1.11	0.46	0.1	0.27	0.25	0.2
	500	0.15	0	1.79	0.15	0.01	0.31	1.15	1.09	0.87	2.17	0.48	0.21	0.48	0.27	0.13
	1000	0.06	0.01	2.39	0.03	0	0.87	1.26	1.25	1.32	2.69	0.67	0.34	0.87	0.26	0.16
	10000	0.01	0	2.8	0.01	0	1.69	1.5	1.48	2.59	2.95	0.51	0.1	0.85	0.14	0.1
f_{16}	50	0.11	0	0.29	0.96	0.93	0.74	0.74	0.71	0.48	0.65	0.29	0	0.34	0.26	0.2
	100	0.1	0.03	0.72	0	0.01	0	0.9	0.91	0.78	1.2	0.46	0.02	0.55	0.25	0.21
	500	0.07	0.02	1.8	0.58	0.57	0.79	1.53	1.52	1.32	2.19	0.5	0	1.21	0.09	0.03
	1000	0	0	2.18	0.27	0.28	0.84	1.21	1.22	1.26	2.49	0.55	0.03	1.24	0	0
	10000	0.04	0	2.97	0.04	0	1.83	1.17	1.15	2.58	3.12	0.59	0.8	1.65	0.04	0

Table 1: Binary logarithms of relative squared Hellinger Risks w.r.t the best method.

Cross-validation using the same set of partitions performs rather poorly. Note that it is particularly bad for the uniform density f_1. Relative performance of **CV** w.r.t. the best method becomes generally worse when sample

size increases. If we compare our penalties and cross-validation for the case of full dynamic programming optimization over a finest regular partition with bin length $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$ ($\mathbf{B}''$, $\mathbf{R}''$,$\mathbf{CV}''$), the picture changes. Overall, the performance of all three methods is not bad, in particular $\mathbf{CV}''$ often outperforms $\mathbf{B}''$ and $\mathbf{R}''$, which behave very similarly. But $\mathbf{CV}''$ performs badly for histogram densities, especially the uniform. Putting a constraint on the minimum bin size causes a problem for all three methods when the density has very sharp peaks (especially f_{10}). For the intermediate case, i.e. using both penalties and cross-validation for a data-driven finest grid but adding a constraint on the bin widths, we see that $\mathbf{B}'$ and $\mathbf{R}'$ share the catastrophic behavior of $\mathbf{B}''$ and $\mathbf{R}''$ at f_{10} without offering a real improvement over $\mathbf{B}$ and $\mathbf{R}$ at the other densities. On the other hand, $\mathbf{CV}'$ is a good compromise between $\mathbf{CV}$ and $\mathbf{CV}''$, but still shows bad behavior for f_1 and f_{10}.

The taut string method **TS** shows a particularly good behavior in terms of Hellinger risk, although it was derived for different aims than performing well w.r.t. a given loss function (Davies and Kovac, 2004), and many questions regarding behavior in a more classical framework remain open. It does not control the number of bins but the modality of the estimate, thereby avoiding overfitting while still being able to chose a large number of bins to give sufficient detail.

Using **AIC** leads to very bad results. It is known to underpenalize even for regular histograms (Birgé and Rozenholc, 2006; Castellan, 1999; Massart, 2007). This tendency becomes even worse when used for irregular histograms, since now there are many partitions with the same number of bins. This

leads to problems similar to those arising in multiple testing. In this case, an additional penalty is needed (Castellan, 1999; Massart, 2007). **BIC** does not aim for a good control of risk but at asymptotically identifying the "smallest true model", if it exists. It shows some good behavior in particular for small sample sizes that deteriorates when samples become larger. Particularly noteworthy is the bad performance for "simple" models like the uniform and the 5 bin regular histogram density f_{13}.

The regular histogram method **BR**, which improves on Akaike's penalization, is the best method for f_3, f_{11} and f_{13}, at least when the sample size is not very small. This shows that the greater flexibility of an irregular histogram over a regular one may be outweighed by the greater difficulty in choosing a good partition, as was already remarked by Birgé and Rozenholc (2006). Regular histograms are inferior for spatially inhomogeneous densities like f_4 and f_{10}.

The simulations show that one can successfully combine the advantages of regular and irregular histograms ($\mathbf{B}^*$ and $\mathbf{R}^*$). The Hellinger risk for $\mathbf{B}^*$ is always within twice the risk of the best method for a given situation. While $\mathbf{R}^*$ is even better in most cases, it shares the bad behaviour of **BR** for f_{10} and $n = 50$. Table 2 shows that for larger sample sizes both $\mathbf{B}^*$ and $\mathbf{R}^*$ almost always choose an irregular histogram for densities where this is advantageous (the spatially inhomogenous densities f_4, f_5, f_{10}, f_{14} and f_{16}) and a regular partition in most of the other cases.

To summarize, we propose a practical method of irregular histogram construction inspired by theoretical works by Barron et al. (1999), Castellan (1999, 2000) and Massart (2007). It can be easily implemented using a

n	method	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
50	$\mathbf{B}^*$	0.08	0.08	0.05	0.75	1.00	0.05	0.14	0.07	0.11	1.00	0.15	0.39	0.06	0.11	0.22	0.51
	$\mathbf{R}^*$	0.02	0.13	0.10	0.68	1.00	0.11	0.27	0.06	0.12	0.00	0.02	0.16	0.03	0.07	0.22	0.43
100	$\mathbf{B}^*$	0.04	0.07	0.04	0.74	1.00	0.11	0.23	0.02	0.19	1.00	0.12	0.43	0.01	0.12	0.24	0.65
	$\mathbf{R}^*$	0.00	0.10	0.10	0.64	1.00	0.24	0.40	0.05	0.36	1.00	0.05	0.28	0.01	0.11	0.33	0.65
500	$\mathbf{B}^*$	0.03	0.00	0.00	0.73	1.00	0.01	0.03	0.00	0.42	1.00	0.00	0.47	0.00	0.69	0.36	0.99
	$\mathbf{R}^*$	0.00	0.00	0.00	0.47	1.00	0.04	0.07	0.00	0.53	1.00	0.00	0.45	0.00	0.75	0.48	0.99
1000	$\mathbf{B}^*$	0.02	0.00	0.00	0.89	1.00	0.00	0.00	0.00	0.58	1.00	0.00	0.41	0.00	0.96	0.49	1.00
	$\mathbf{R}^*$	0.01	0.00	0.00	0.63	1.00	0.01	0.00	0.00	0.64	1.00	0.00	0.41	0.00	0.98	0.57	1.00
5000	$\mathbf{B}^*$	0.01	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.65	1.00	0.00	0.26	0.00	1.00	0.43	1.00
	$\mathbf{R}^*$	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.70	1.00	0.00	0.28	0.00	1.00	0.47	1.00
10000	$\mathbf{B}^*$	0.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.66	1.00	0.00	0.18	0.00	1.00	0.35	1.00
	$\mathbf{R}^*$	0.00	0.00	0.00	1.00	0.98	0.00	0.00	0.00	0.71	1.00	0.00	0.20	0.00	1.00	0.39	1.00

Table 2: Frequency of choosing an irregular partition.

dynamic programming algorithm and it performs well for a wide range of different densities and sample sizes, even for some cases not covered by the underlying theory. Performance is shown to be improved when combined with the regular histogram approach proposed in Birgé and Rozenholc (2006). All procedures proposed here are available in the `R`-package `histogram` (Mildenberger et al., 2009a).

Acknowledgments

This work has been supported in part by the Collaborative Research Centers "Reduction of Complexity in Multivariate Data Structures" (SFB 475) and "Statistical Modelling of Nonlinear Dynamic Processes" (SFB 823) of the German Research Foundation (DFG). The authors also wish to thank Henrike Weinert for discussions and programming in earlier stages of the work as well as two anonymous referees for giving comments that greatly improved the paper.

References

Akaike, H., 1973. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

Barron, A., Birgé, L., Massart, P., 1999. Risk bounds for model selection via penalization. Probability Theory and Related Fields 113, 301-413.

Berlinet, A., Devroye, L., 1994. A comparison of kernel density estimates. Publications de l'Institut de Statistique de l'Universite de Paris 38, 3-59.

Birgé, L., Rozenholc, Y., 2006. How many bins should be put in a regular histogram? ESAIM: Probability and Statistics 10, 24-45.

Blanchard, G., Schäfer, C., Rozenholc, Y., Müller, K.-R., 2007. Optimal dyadic decision trees. Machine Learning 66, 209-241.

Castellan, G., 1999. Modified Akaike's criterion for histogram density estimation. Technical Report 99.61, Université de Paris-Sud.

Castellan, G., 2000. Sélection d'histogrammes à l'aide d'un critère de type Akaike. Comptes rendus de l'Académie des sciences Paris 330, Série I, 729-732.

Catoni, O., 2002. Data compression and adaptive histograms, in: Cucker, F., Rojas J.M. (Eds.), Foundations of Computational Mathematics, Proceedings of the Smalefest 2000, World Scientific, Singapore, pp. 35-60.

Celisse, A., Robin, S., 2008. Nonparametric density estimation by exact leave-p-out cross-validation. Computational Statistics and Data Analysis 52, 2350-2368.

Chen, X.R., Zhao, L.C., 1987. Almost sure L_1-norm convergence for data-based histogram density estimators. Journal of Multivariate Analysis 21, 179-188.

Comte, F., Rozenholc, Y., 2004. A new algorithm for fixed design regression and denoising. Annals of the Institute of Statistical Mathematics 56, 449-473.

Davies, P. L., Gather, U., Nordman, D. J., Weinert, H., 2009. A comparison of automatic histogram constructions. ESAIM: Probability and Statistics 13, 181-196.

Davies, P. L., Kovac, A., 2004. Densities, spectral densities and modality. The Annals of Statistics 32, 1093-1136.

Davies, P.L., Kovac, A., 2008. ftnonpar: Features and strings for nonparametric regression. R package version 0.1-83.

Devroye, L., Györfi, L., 1985. Nonparametric density estimation: the L_1 view. Wiley, New York.

Devroye, L., Lugosi, G.. 2004. Bin width selection in multivariate histograms by the combinatorial method, Test 13, 129-145.

Engel, J., 1997. The multiresolution histogram. Metrika 46, 41-57.

Hartigan, J.A., 1996. Bayesian histograms, in: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics 5, Oxford University Press, Oxford, pp. 211-222.

Kanazawa, Y., 1988. An optimal variable cell histogram. Communications in Statistics - Theory and Methods 17, 1401-1422.

Kanazawa, Y., 1992. An optimal variable cell histogram based on the sample spacings. The Annals of Statistics 20,219-304.

Klemelä, J., 2007. Density estimation with stagewise optimization of the empirical risk. Machine Learning 67, 169-195.

Klemelä, J., 2009. Multivariate histograms with data-dependent partitions. Statistica Sinica 19, 159-176.

Kogure, A., 1986. Optimal cells for a histogram. PhD thesis, Yale University.

Kogure, A., 1987. Asymptotically optimal cells for a histogram. The Annals of Statistics 15, 1023-1030.

Kontkanen, P., Myllymäki, P., 2007. MDL histogram density estimation. In: Meila M., Shen S. (Eds.), Proc. 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007), Puerto Rico, March 2007. http://www.stat.umn.edu/∼aistat/proceedings/start.htm

Lugosi, G., Nobel, A., 1996. Consistency of data-driven histogram methods for density estimation and classification. The Annals of Statistics 24, 687-706.

Massart, P., 2007. Concentration inequalities and model selection. Lecture Notes in Mathematics Vol. 1896, Springer, New York.

Mildenberger, T., Rozenholc, Y., Zasada, D., 2009a. histogram: Construction of regular and irregular histograms with different options for automatic choice of bins. R package version 0.0-23.

Mildenberger, T., Weinert, H., Tiemeyer, S., 2009b. benchden: 28 benchmark densities from Berlinet/Devroye (1994). R package version 1.0.3.

Rissanen, J., Speed, T. P., Yu, B., 1992. Density estimation by stochastic complexity. IEEE Transactions on Information Theory 38, 315-323.

Rozenholc, Y., Mildenberger, T., Gather, U., 2009. Combining regular and irregular histograms by penalized likelihood. Discussion Paper 31/2009, SFB 823, Technische Universität Dortmund.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461-464.

Zhao, L.C, Krishnaiah, P.R., Chen, X.R., 1988. Almost sure L_r-norm convergence for data-based histogram estimates. Theory of Probability and its Applications 35, 396-403.