Rozenholc, Yves; Mildenberger, Thoralf; Gather, Ursula

Working Paper

Constructing irregular histograms by penalized likelihood

Technical Report, No. 2009,04

Provided in Cooperation with:
Collaborative Research Center 'Reduction of Complexity in Multivariate Data Structures' (SFB 475), University of Dortmund

Suggested Citation: Rozenholc, Yves; Mildenberger, Thoralf; Gather, Ursula (2009) : Constructing irregular histograms by penalized likelihood, Technical Report, No. 2009,04, Technische Universität Dortmund, Sonderforschungsbereich 475 - Komplexitätsreduktion in Multivariaten Datenstrukturen, Dortmund

This Version is available at:
https://hdl.handle.net/10419/41047

Constructing Irregular Histograms by Penalized Likelihood

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

[a] *UFR de Mathmatiques 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

We propose a fully automatic procedure for the construction of irregular histograms. For a given number of bins, the maximum likelihood histogram is known to be the result of a dynamic programming algorithm. To choose the number of bins, we propose two different penalties motivated by recent work in model selection by Castellan [6] and Massart [26]. We give a complete description of the algorithm and a proper tuning of the penalties. Finally, we compare our procedure to other existing proposals for a wide range of different densities and sample sizes.

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

1. Introduction

A histogram is a piecewise constant probability density. We first introduce some notation. 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 their order statistics by $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Given a partition $\mathcal{I}$ of a compact interval $K \subset \mathbb{R}$ into D subintervals, consider all histograms piecewise constant on $\mathcal{I}$ and zero outside $\mathcal{I}$, i.e. functions of the form

$$\hat{f}(x) = \sum_{j=1}^{D} h_j \mathbb{1}_{I_j}(x)$$

where $\mathbb{1}_A$ denotes the indicator function of A and where $h_1, \ldots, h_D \geq 0$ are such that the integral of $\hat{f}$ is 1; $\hat{f}$ can be regarded as an estimate of f. If K contains $[x_{(1)}, x_{(n)}]$, among all histograms associated to the partition $\mathcal{I}$, the *Maximum Likelihood Histogram*

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

(ML histogram) is given by the histogram $\hat{f}_{\mathcal{I}}$ defined by

$$\hat{f}_{\mathcal{I}} := \frac{1}{n} \sum_{j=1}^{D} \frac{N_j}{|I_j|} \mathbb{1}_{I_j}, \tag{1}$$

with $N_j = \sum_{i=1}^{n} \mathbb{1}_{I_j}(x_i)$ and $|I_j|$ the length of the interval I_j. Its loglikelihood is

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

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 an irregular 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 with respect to the following loss functions:

- Squared Hellinger distance

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

  which has been normalized such that its maximum value is 1.

- Powers of the L_p-norms (for $p = 1$ and 2) defined by

$$d_p := \|f - g\|_p^p = \int |f(t) - g(t)|^p dt. \tag{3}$$

The L_2 distance is widely used mainly for its mathematical tractability. It is often possible to derive explicit expressions for the L_2 risk at least asymptotically, cf. Kogure [23]. However, as argued by Devroye and Györfi [15], ch. 1, the L_1 distance can be considered more natural in the context of density estimation because – unlike the L_2 distance – it

is defined for all densities and it has desirable invariance properties. We mainly focus on the Hellinger distance for several reasons: it is also defined for any two densities, it has important invariance properties and the results of Castellan [6, 7] are derived for the corresponding risk. Another widely used loss function is the Kullback-Leibler distance

$$d_{KL}(f,g) := \int \log\left(\frac{f(t)}{g(t)}\right) f(t)dt$$

which is not suitable in histogram density estimation since it is infinite whenever the estimated density is zero on an interval where the true distribution has positive mass. Hence it is excluded from consideration. For a detailed discussion on the choice of loss functions in histogram density estimation, see section 2.2. in Birgé and Rozenholc [4] and the references given there.

Given the sample, the histogram $\hat{f}_{\mathcal{I}}$ depends only on the chosen partition $\mathcal{I} = (I_1, \ldots, I_D)$. The values on the intervals of the partition are fixed, namely equal to the relative frequencies divided by the bin widths. The crucial point is thus choosing the partition. A naïve comparison of the likelihood of histograms for partitions with different numbers of bins is misleading since partitions with too many bins will result in a large likelihood without yielding a sensible estimate of f. 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 [4] and Davies, Gather, Nordman and Weinert [12] and the references given there.

Several methods have been developed to choose a good irregular histogram. Kogure [23] gives asymptotic results for the optimal choice of bins. His approach is based on using blocks of equisized bins, and the dependence on tuning parameters is explored via simulations in his PhD thesis [22]. It does not result in a fully automatic procedure. Kanazawa [19] 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 [20] derives the asymptotically optimal choice of the number of bins. Unfortunately this result involves the first and second derivatives of the unknown density, which leads to a construction that cannot be applied from a practical point of view. Celisse and Robin [9] 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 [17] is based on a tree of dyadic partitions to control the L_2-error. The 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, Schäfer, Rozenholc and Müller [5]) or a greedy search on a full set of partitions (Klemelä [21]) to deal with computational problems that

do not occur in the univariate case. Theoretical results on conditions for consistency of histogram estimates with data-driven and possibly irregular partitions are derived in Chen and Zhao [10], Zhao, Krishnaiah and Chen [30], Lugosi and Nobel [25]. Devroye and Lugosi [16] give a construction of histograms where bin widths are allowed to vary according to a pre-specified function.

Hartigan [18] 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, Speed and Yu [28] give a construction based on the Minimum Description Length (MDL) paradigm, which leads to a penalized likelihood estimator. 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 makes computing a histogram computationally expensive. A more recent proposal by Kontkanen and Myllymäki [24] is also based on the MDL principle; it also involves a discretization which results in the estimate not being a proper density. Catoni [8] suggests a multi-stage procedure that computes a density estimate by aggregating histograms which is also based on coding ideas.

The taut string procedure introduced by Davies and Kovac [13] can also be used to generate an irregular histogram as described in Davies, Gather, Nordman and Weinert [12]. Regularization is performed not by controlling the number of bins but by controlling the modality of the estimate. The stated aim of the authors is not to minimize some risk but to find an estimate of the density that has minimum number of modes that could have generated the data, where the latter is formalized by a criterion based on differences of Kuiper metrics between the empirical and the estimated distribution. 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. 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. Let us emphasize that although the partition is chosen without reference to maximum likelihood, the histogram constructed in this way fulfils definition (1). The main tuning parameter is the tube width, and an automatic choice is suggested by the authors. The procedure has shown a particularly good behavior also w.r.t. classical loss functions (Davies, Gather, Nordman and Weinert [12]), and therefore is compared with our method in our simulations.

Here we will focus on automatic construction of irregular histograms using penalized likelihood maximization techniques. For a good data-driven choice of the estimated 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 [1], 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 introduced by Schwarz [29]. It is constructed to consistently estimate the smallest true model order, which in histogram density estimation is infinite unless the true density is piecewise constant. In practice, criteria like AIC and BIC [1, 29] are routinely applied in many different statistical models, often without refer-

ence 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 [26] 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 due to Barron, Birgé and Massart [2], Castellan [6, 7] and Massart [26].

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 choice of the penalty. Section 4 gives a detailed description of the proposed procedure for irregular histograms. In section 5, we comment on the empirical evaluation of the risks under consideration. Section 6 gives the results of a simulation study and conclusions.

2. Penalized likelihood construction of histograms

Constructing an irregular histogram by penalized likelihood means maximizing w.r.t. partitions $\mathcal{I} = (I_1, \ldots, I_{|\mathcal{I}|})$ of $[x_{(1)}, x_{(n)}]$:

$$L(\hat{f}_{\mathcal{I}}, x_1, \ldots, x_n) - \mathrm{pen}_n(\mathcal{I}) \tag{4}$$

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, Birgé and Massart [2], Castellan [6, 7] and Massart [26].

Optimizing w.r.t. the partition $\mathcal{I}$ with $|\mathcal{I}|$ fixed in (4) leaves us with a continuous optimization problem. Without further restrictions, for $|\mathcal{I}| \geq 2$ the likelihood is unbounded. The partition

$$\{[x_{(1)}, x_{(1)} + \eta), [x_{(1)} + \eta, x_{(n)}]\}$$

leads to a log-likelihood equal to

$$-n \log(n) - \log(\eta) - (n-1)(\log(n-1) + \log(x_{(n)} - x_{(1)} - \eta))$$

which can be arbitrarily large when η goes to 0.

One possibility is to restrict to all partitions which are built with endpoints on the observations; the optimization problem (4) can then be solved using a dynamic programming algorithm first used for histogram construction by Kanazawa [19]. More details are given in Section 4.

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

$$\mathrm{pen}_n^{(1)}(\mathcal{I}) = c \log \binom{n-1}{D-1} + \alpha(D-1) + \varepsilon_{c,\alpha}^{(1)}(D) \tag{5}$$

$$\mathrm{pen}_n^{(2)}(\mathcal{I}) = c \log \binom{n-1}{D-1} + \frac{\alpha}{n} \sum_{j=1}^{D} \frac{N_j}{|I_j|} + \varepsilon^{(2)}(D). \tag{6}$$

where

$$\varepsilon^{(1)}_{c,\alpha}(D) = c\,k\,\log D + 2\sqrt{c\alpha(D-1)(\log\binom{n-1}{D-1} + k\log D)}\,, \tag{7}$$

$$\varepsilon^{(2)}(D) = \log^{2.5} D. \tag{8}$$

The precise choices for c and α obtained by simulations are described in Section 3.

We now give arguments to explain the origins of these penalties. The penalty defined by (5) is derived from Theorem 3.2 in Castellan [6], which is also stated as Theorem 7.9 in Massart [26], p. 232 and from eq. (7.32) in Theorem 7.7 in Massart [26], p.219. The penalty defined by (6) comes from eq. (7.33) in Theorem 7.7 in Massart [26]. From the penalty form in Theorem 7.9 in Massart [26] we derive $\varepsilon^{(1)}$:

$$\mathrm{pen}_n(\mathcal{I}) = c_1(\sqrt{D-1} + \sqrt{c_2 x_{\mathcal{I}}})^2 \tag{9}$$

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

$$\sum_D \sum_{|\mathcal{I}|=D} e^{-x_{\mathcal{I}}} \leq \Sigma \tag{10}$$

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 x_D to every partition $\mathcal{I}$ with $|\mathcal{I}| = D$ such that

$$\sum_D \binom{n-1}{D-1} e^{-x_D} \leq \Sigma.$$

To achieve this, we set

$$x_D = \log\binom{n-1}{D-1} + \varepsilon(D)$$

Then (10) becomes

$$\sum_D e^{-\varepsilon(D)} \leq \Sigma.$$

Choosing $\varepsilon(D)$ of the form $k\log D$ with $k > 1$ ensures that the sum is converging and that Σ is finite. Finally for $k > 1$ we have

$$x_D = \log\binom{n-1}{D-1} + k\log D.$$

Substitution into (9) gives

$$\begin{aligned}\mathrm{pen}_n(\mathcal{I}) = c_1 \Bigg(& D - 1 + c_2 \left(\log\binom{n-1}{D-1} + k\log D\right) \\ & + 2\sqrt{c_2(D-1)\left(\log\binom{n-1}{D-1} + k\log D\right)}\Bigg).\end{aligned} \tag{11}$$

Let us emphasize that Theorem 7.9 in Massart [26], p. 232 states $c_1 > 1/2$ 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 (7).

We want now to use Theorem 7.7 in Massart [26], p. 219 to justify the penalty in (6). The orthonormal basis considered in this theorem for a given partition $\mathcal{I}$ consists of all $\mathbb{1}_I/\sqrt{|I|}$ 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 loglikelihood, 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 [26], p. 219, following the same derivation for $\varepsilon^{(1)}$, we find the penalty in (11) 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 [26]. 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 (6) we start from the penalty defined in (7.33) in Massart [26]:

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

Following the same derivations as for the penalty (9), 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 term in the square root above breaks the possibility to use dynamical programming to compute the maximum of the penalized loglikelihood defined in (4). To avoid this problem we propose, following penalty forms proposed in Birgé and Rozenholc [4] and Comte and Rozenholc [11], to replace the remainder term

$$2k\log D + 2\sqrt{2\widehat{V}_{\mathcal{I}}\left(\log\binom{n-1}{D-1} + k\log D\right)}$$

by a power of $\log D$. We have tried several values of the power to finally conclude that Formula (8) leads to a good choice.

3. Choice of the Penalty

Using histograms with endpoints of the partitions placed on the observations as described later in Section 4, we ran empirical risk estimation in order to fix our penalty using the losses defined by (2) and (3) for $p = 1$ and 2. We focused on the Hellinger risk to obtain good choices of the penalties, but the behavior w.r.t. L_1 and L_2 losses was not very different. Since no single penalty is best in all cases, the calibration of a penalty always leads to some compromise. We describe in the following what we consider to be a good proposal. We start with the random penalty as it is simpler.

3.1. Random Penalty

In formula (6) we ran risk evaluation experiments using all combinations with $c \in \{0.5, 1, 2\}$ and $\alpha \in \{0.5, 1\}$. Let us emphasize that $c = 2$ and $\alpha = 1$ corresponds to Formula (7.33) in Massart [26] up to our choice of $\varepsilon^{(2)}$ defined in (8). From our point of view, the most satisfactory choice is $c = 1$ and $\alpha = 0.5$.

3.2. Deterministic Penalty

In formula (5) we have chosen:

- $c = 2(\alpha + 1)$ and $\alpha \in \{0.5, 1\}$ following Theorem 7.9 in Massart [26].
- $c = 2$ and $\alpha = 1$ following Theorem 7.7 eq. (7.32) in Massart [26] with $M = 1$ and $\varepsilon = 0$.
- $c = 1$ and $\alpha \in \{0.5, 1\}$.

From these experiments, the most satisfactory choice we have found is $c = 2$ and $\alpha = 1$. In this deterministic penalty framework, we also ran experiments replacing $\varepsilon^{(1)}_{c,\alpha}$ by $\varepsilon^{(2)}$. In this case, we have found that the most satisfactory choice is $c = 1$ and $\alpha = 1$, and this choice is even better than $\varepsilon^{(1)}_{2,1}$.

To conclude this section, we remark that the results are very close. Only for the trimodal uniform density, we have found differences in favor of the deterministic penalty. For all other densities, the absolute values of the relative differences $\left|\frac{\widehat{R}^{\mathrm{R}}_n - \widehat{R}^{\mathrm{D}}_n}{\widehat{R}^{\mathrm{D}}_n}\right|$ of the risks are less than 0.162.

4. Construction of the Penalized Maximum Likelihood Histogram

We maximize (4) w.r.t. 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 $1 < k_1 < \ldots < k_{D_{\max}-1} < n$. Let us write this partition as

$$\mathcal{I}_{\max} = (I^0_1, \ldots, I^0_{D_{\max}}),$$

where $I_d^0 = (t_{d-1}, t_d]$ for $d = 1$ to $D_{\max}$ and where $t_0 = x_{(1)} - \text{eps}$, $t_{D_{\max}} = x_{(n)}$ and $t_d = x_{(k_d)}$ for $0 < d < D_{\max}$. Here eps represents the machine precision and is used only to help the use of left-open, right-closed intervals. Our aim is to build a sub-partition $\mathcal{I}$ of $\mathcal{I}_{\max}$ which maximizes (4). This problem is solved in polynomial time by a dynamic programming (DP) algorithm as used in Kanazawa [19] and Comte and Rozenholc [11]. We briefly describe the algorithm in our context of penalized histograms. Let us assume that (4) can be rewritten (up to the knowledge of the sample) as $\Phi^0(\mathcal{I}) + \Psi(D, n)$, where Φ^0 is an additive function with respect to the partition in the sense that

$$\Phi^0(\mathcal{I}) = \Phi(I_1) + \ldots + \Phi(I_D) \text{ if } \mathcal{I} = (I_1, \ldots, I_D).$$

In our case, $\Phi(I)$ depends only on the number N_I of data fallen in interval I and on its length $|I|$. More precisely for a penalty of the form (5)

$$\Phi(I) = N_I \log \frac{N_I}{n|I|} \tag{12}$$

and for a penalty of the form (6) we have

$$\Phi(I) = N_I \log \frac{N_I}{n|I|} - \frac{\alpha}{n} \frac{N_I}{|I|},$$

while $\Psi(D, n) = \text{pen}_n^{(1)}(\mathcal{I})$ in the deterministic case and

$$\Psi(D, n) = c \log \binom{n-1}{D-1} + \varepsilon^{(2)}(D),$$

in the random case.

We denote by $p_1(i, j) = \Phi((t_i, t_j])$ and $p_1(j) := p_1(0, j)$. Finally, let us define $i_1(j) = 0$. Assume that we have already computed all $p_1(i, j)$ for $0 \leq i < j \leq D_{max}$ (which needs $O(D_{max}^2)$ operations). The dynamic programming algorithm works as follows:

- `For` $D = 2 \ldots D_{\max}$
  - `For` $j = D \ldots D_{\max}$,
    - $i_D(j) = \arg_i \max_{D-1 \leq i < j} [p_{D-1}(i) + p_1(i, j)]$;
    - $p_D(j) = p_{D-1}(i_D(j)) + p_1(i_D(j), j)$

$p_D(D_{\max})$ is the maximum of $\Phi^0(\mathcal{I})$ for all sub-partitions $\mathcal{I}$ - of our finest partition $\mathcal{I}_{\max}$ - with D bins. The partition which achieves the maximum of $\Phi^0(\mathcal{I}) + \Psi(D, n)$ may be built in the following way:

- `Compute` $\widehat{D} = \arg_D \max_{1 \leq D \leq D_{\max}} p_D(D_{\max}) + \Psi(D, n)$.
- `Fix` $L = D_{max}$

- `For` $j = D, \ldots, 1$, `grow a vector` $L := [L, i_j(L(\text{last}))]$
- `Reverse the order of the vector` L

The vector L defines the index of the t_j's which are the endpoints of the best partition in the sense of (4). The notation $L(\text{last})$ denotes the last coordinate of the vector L and $[L, u]$ denotes concatenation of the vector L with u.
The computation of $i_D(j) = \arg_i \max_{D-1 \le i < j}[p_{D-1}(i) + p_1(i, j)]$ requires $O(j - D + 1)$ operations and the total complexity of this algorithm is of order $D_{\max}^3$. Hence the total number of operations may be of order n if we start from a finest partition with $D_{\max}$ of order $n^{1/3}$ or $n^{1/3} \log n$. We propose to use a greedy algorithm in order to build this finest partition. Let us call $\mathcal{E}(\mathcal{I})$ the set of endpoints of partition $\mathcal{I}$. Starting with the partition $\mathcal{I}_0 = ([x_{(1)}, x_{(n)}])$, we grow a sequence of partitions $\mathcal{I}_D$ satisfying :

$$\mathcal{I}_{D+1} = \arg\max \Phi^0(\mathcal{I}),$$

where the maximum is taken over all partitions $\mathcal{I}$ with $\mathcal{E}(\mathcal{I}) = \mathcal{E}(\mathcal{I}_D) \cup \{t\}$ with t in $\{x_1, \ldots, x_n\} \setminus \mathcal{E}(\mathcal{I}_D)$. For both penalty forms, we use a greedy maximization of the likelihood to obtain this partition, i.e. we always use Φ as in (12).

Let us remark that the theoretical results by Castellan [6, 7] and Massart [26], ch. 7, are derived for the case of a finest regular grid with bin sizes not smaller than a constant times $\log^2(n)/n$. In particular, the set of partitions is fixed beforehand and may depend on n but not on the sample. This also means that that no bins are possible that are shorter than a constant times $\log^2(n)/n$. However, we found that, in practice, we can improve performance drastically for densities by using a data-dependent finest grid imposing no restrictions on the smallest bins without loosing much at other densities. More comments on this are given in section 6.

5. Risk evaluation

The risks of the procedures are evaluated empirically by means of simulations. For each density f and each sample size n, N samples $x^{(j)} := (x_1^{(j)}, \ldots, x_n^{(j)})$, $j = 1, \ldots, N$ are generated and the loss functions $d = d_H, d_1, d_2$ are evaluated for every histogram procedure $\hat{f}$. We estimate the risks $R_n(f, \hat{f}, d)$ by

$$\widehat{R}_n(f, \hat{f}, d) := \sum_{j=1}^{N} d(f, \hat{f}(x_1^{(j)}, \ldots, x_n^{(j)})).$$

We now describe how we computed our loss functions (2) and (3) to obtain empirical risk evaluation. To estimate the risks, we evaluate the losses $d(f, \hat{f}(x_1^{(j)}, \ldots, x_n^{(j)}))$ for every simulation run j by numerical integration. First note that the integrals appearing in (2) and (3) are all of the form

$$\int \delta(t)dt := \int \tilde{\delta}(f(t), g(t))dt$$

for continuous functions $\tilde{\delta}$. Care has to be taken of discontinuities in both the true densities f and the histogram estimates $\hat{f}$ and furthermore the bilogarithmic peak density has infinite peaks. For given f and $\hat{f}$, let $\tau_1 < \cdots < \tau_{L-1}$ denote the points where f or $\hat{f}$ is discontinuous or infinite. Defining the intervals $J_0 := (-\infty, \tau_1)$, $J_l := (\tau_l, \tau_{l+1})$, $l = 1, \ldots, L-1$, $J_L := (\tau_L, \infty)$, we split up the integrals into sums of integrals over open intervals where δ is continuous:

$$\int_{\mathbb{R}} \delta(t)dt := \sum_{l=0}^{L} \int_{J_l} \delta(t)dt.$$

Note that we use open intervals to allow both f and $\hat{f}$ to take any (possibly infinite) value in the point $\tau_1, \ldots, \tau_l$. To evaluate the integrals on $J = J_1, \ldots, J_{L-1}$ we use the trapez rule

$$\int_{J_l} \delta(t)dt \approx (\kappa_K^l - \kappa_1^l)\left(\frac{1}{2}\delta(\kappa_1^l) + \delta(\kappa_2^l) + \cdots + \delta(\kappa_{K-1}^l) + \frac{1}{2}\delta(\kappa_K^l)\right)$$

for equispaced grid points $\kappa_1^l = \tau_l + \varepsilon$, $\kappa_K^l = \tau_{l+1} - \varepsilon$ and $\kappa_\nu^l = \kappa_1^l + (\nu - 1)h$ for $\nu = 2, \ldots, K-1$ with $h = \frac{\tau_{l+1} - \tau_l - 2\varepsilon}{K-1}$. We set $\varepsilon = 10^{-11}$ to integrate over open intervals.

Note that on the unbounded intervals J_0 and J_L for $d = d_H$ and $d = d_1$ we have $\int_{J_l} \delta(t)dt = \int_{J_l} f(t)dt$ since $\hat{f}$ is zero. For d_2 we replace $\pm\infty$ in the definition of J_0 and J_L by the upper and lower 10^{-10}-quantiles of f and integrate numerically as on the other intervals, in case the support of f is unbounded. Otherwise, the integrals over J_0 and J_L are zero.

6. Simulation Study and Conclusions

In order to tune the constants in the penalties given in Section 3 and to assess the performance of the penalized likelihood histogram defined as the maximizer of (4) with penalty defined by (5) or (6), we conducted a simulation study involving empirical risk estimation with respect to the losses (2) and (3) (for p=1,2). The choices we arrive at are given in section 3. Then we compare our choices for the penalized maximum likelihood to other available methods in a separate simulation study using the same densities.

Performance of the methods is compared on 12 of the 28 test-bed densities introduced by Berlinet and Devroye (1994) and implemented in the `R`-package `benchden` [27]. We used densities 1 (uniform), 4 (double exponential), 11 (normal), 12 (lognormal), 21-24 (mixtures of normals) and 25-28 (various other multimodal densities). We denote these by $f_1, \ldots, f_{12}$. We also added 4 histogram densities:

- 5 bin regular histogram:

$$\begin{aligned} f_{13}(x) \;:=\; & 0.15u_{[0,0.2]}(x) + 0.35u_{(0.2,0.4]}(x) + 0.2u_{(0.4,0.6]}(x) \\ & + 0.1u_{(0.6,0.8]}(x) + 0.2u_{(0.8,1.0]}(x) \end{aligned}$$

- 5 bin irregular histogram:

$$\begin{aligned} f_{14}(x) &:= 0.15u_{[0,0.13]}(x) + 0.35u_{(0.13,0.34]}(x) + 0.2u_{(0.34,0.61]}(x) \\ &\quad +0.1u_{(0.61,0.65]}(x) + 0.2u_{(0.65,1.0]}(x) \end{aligned}$$

- 10 bin regular histogram:

$$\begin{aligned} f_{15}(x) &:= 0.01u_{[0,0.1]}(x) + 0.18u_{(0.1,0.2]}(x) + 0.16u_{(0.2,0.3]}(x) \\ &\quad +0.07u_{(0.3,0.4]}(x) + 0.06u_{(0.4,0.5]}(x) + 0.01u_{(0.5,0.6]}(x) \\ &\quad +0.06u_{(0.6,0.7]}(x) + 0.37u_{(0.7,0.8]}(x) + 0.06u_{(0.8,0.9]}(x) \\ &\quad +0.02u_{(0.9,1.0]}(x) \end{aligned}$$

- 10 bin irregular histogram:

$$\begin{aligned} f_{16}(x) &:= 0.01u_{[0,0.02]}(x) + 0.18u_{(0.02,0.07]}(x) + 0.16u_{(0.07,0.14]}(x) \\ &\quad +0.07u_{(0.14,0.44]}(x) + 0.06u_{(0.44,0.53]}(x) + 0.01u_{(0.53,0.56]}(x) \\ &\quad +0.06u_{(0.56,0.67]}(x) + 0.37u_{(0.67,0.77]}(x) + 0.06u_{(0.77,0.91]}(x) \\ &\quad +0.02u_{(0.91,1.0]}(x) \end{aligned}$$

where $u_I := |I|^{-1}\mathbb{1}_I$ denotes the uniform density on an interval I. All densities are depicted in Figure 1. Note that Castellan's main theorem 3.2 in [6] does not apply to all densities considered here, since she assumes e.g. that the density is bounded away from zero. We include a wide range of densities in order to explore the behavior of the procedure also in cases not covered by theory. The sample sizes are 50,100,500,1000,5000 and 10000. We used 500 replications for each scenario and estimated the resulting risks as described in section 5 using $\kappa = 5000$.

The methods compared in the simulations are (abbreviations in parentheses correspond to column titles in tables 2 and 3 in the appendix A):

- Penalized maximum likelihood using deterministic penalty (5) with $c = 1$ and $\alpha = 1$. Maximiziation is performed over a data-driven finest grid as described in section 4 without restrictions on the minimum bin width. **(D)**

- Penalized maximum likelihood using random penalty (6) with $c = 1$ and $\alpha = 0.5$. Maximiziation is performed over a data-driven finest grid as described in section 4 without restrictions on the minimum bin width. **(R)**

- Leave-one-out cross-validation using formula (11) given in [9] with the same set of partitions as for our two proposals **(D)** and **(R)**. We also tried formula (12) of [9] for different values of p without finding a big difference. **(CV)**

- Methods 1-3 using the same data-driven grid but with the additional constraint that the minimum allowed bin length is $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$. **(Dc)**, **(Rc)**,**(CVc)**

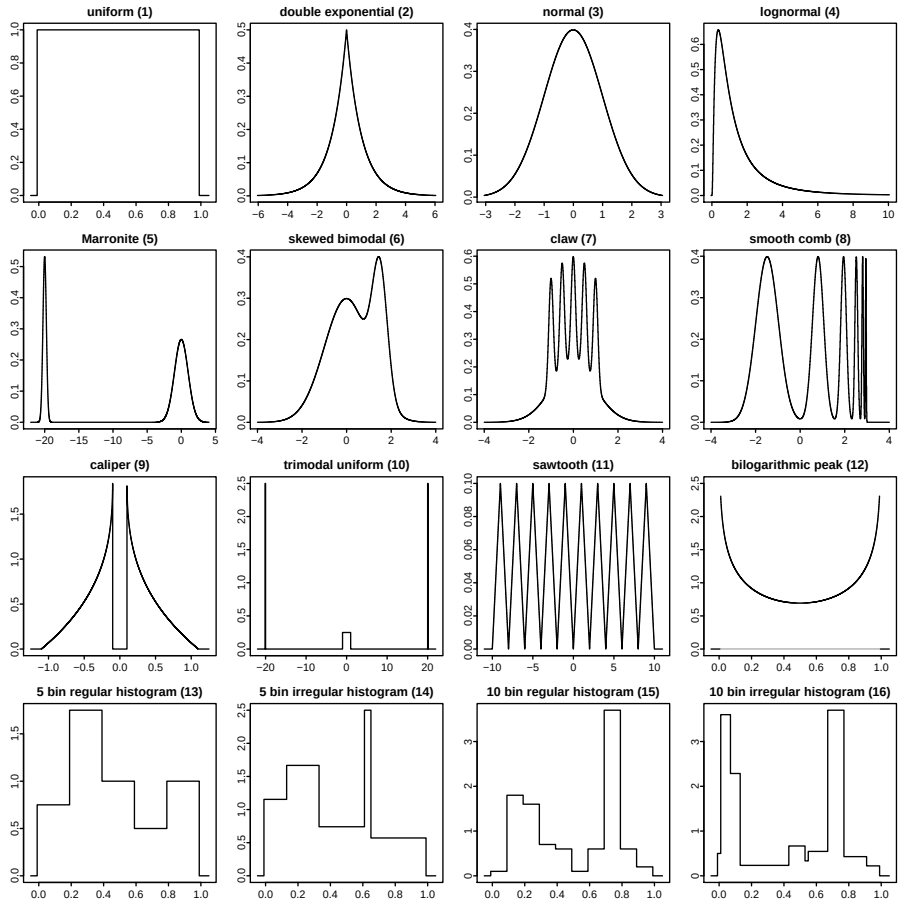

Figure 1: The densities used in the simulation study

- Methods 1-3 using a full optimization over a finest regular partition with bin width $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$. This is the grid considered in [6], except that we slightly relax her $\log^2(n)$ to $\log^{1.5}(n)$. **(Dr)**, **(Rr)**,**(CVr)**

- Penalized maximum likelihood using Akaike's Information Criterion introduced by Akaike [1]. The penalty is $\mathrm{pen}_n^{\mathrm{AIC}}(D) = (D-1)$. **(AIC)**

- Penalized maximum likelihood using the Bayesian Information Criterion introduced by Schwarz [29]. The penalty is
$\mathrm{pen}_n^{\mathrm{BIC}}(D) = 0.5 \log(n)(D-1)$. **(BIC)**

- The taut-string method introduced by Davies and Kovac [13]. We use the function `pmden()` implemented in the R-package `ftnonpar` [14] 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. **(TS)**

- Regular histogram construction due to Birgé and Rozenholc [4]. The penalty is $\mathrm{pen}_n^{\mathrm{BR}}(D) = D + \log(D)^{2.5}$, where the loglikelihood is maximized over all regular partitions with $1, \ldots, [n/\log n]$ bins. We use this as a reference method to highlight advantages and disadvantages of using different irregular histogram methods over using a well-tuned regular histogram. **(BR)**

For the discussion of the results, we focus on squared Hellinger risk, but the results for L_1 and L_2 are not very different. Table 1 gives the empirical risk results (multiplied by 100) for the two methods that showed the overall best performance: maximum penalized likelihood using a data driven grid with the random penalty (6) and no constraints on minimum bin width and the taut string method. The table shows no obvious winner between those two.

Table 2 in the appendix shows the dyadic logarithms of relative risks w.r.t. the best method for any given n and density: $\log_2(\widehat{R}_n^{\mathrm{method}}/\widehat{R}_n^{\mathrm{best}})$ for all 13 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. Table 3 in the appendix shows the modes of the number of bins chosen for all methods as well as the corresponding frequencies with which this number was chosen.

n	method	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8
50	**(R)**	2.09	6.30	5.42	5.93	8.40	4.73	6.14	10.63
	(TS)	2.13	5.44	5.75	4.18	7.35	5.44	7.21	10.52
100	**(R)**	1.02	4.55	3.42	3.97	5.97	2.97	3.81	7.40
	(TS)	1.07	3.23	3.09	2.62	4.40	3.31	3.97	6.83
500	**(R)**	0.20	1.80	1.38	1.71	2.21	1.33	2.31	2.98
	(TS)	0.22	1.14	1.00	1.12	1.29	1.05	1.48	2.38
1000	**(R)**	0.10	1.20	0.92	1.12	1.51	0.89	1.56	1.95
	(TS)	0.11	0.74	0.65	0.84	0.80	0.66	0.85	1.49
5000	**(R)**	0.02	0.46	0.35	0.46	0.57	0.34	0.58	0.75
	(TS)	0.02	0.28	0.25	0.26	0.28	0.19	0.26	0.45
10000	**(R)**	0.01	0.30	0.23	0.30	0.37	0.22	0.39	0.50
	(TS)	0.01	0.18	0.17	0.18	0.18	0.12	0.16	0.26

n	method	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
50	**(R)**	11.66	21.55	7.49	3.79	4.27	4.76	7.91	6.65
	(TS)	10.49	9.65	8.46	3.58	4.25	4.96	8.56	6.65
100	**(R)**	4.86	4.16	6.46	2.62	3.06	3.22	4.62	3.86
	(TS)	4.50	5.21	7.64	2.59	3.06	3.37	4.36	3.81
500	**(R)**	1.67	0.83	4.15	0.89	0.90	0.74	1.17	1.18
	(TS)	1.19	1.18	2.93	0.71	0.77	0.77	1.35	1.16
1000	**(R)**	1.07	0.42	2.55	0.62	0.34	0.37	0.54	0.63
	(TS)	0.72	0.61	1.81	0.44	0.38	0.35	0.68	0.64
5000	**(R)**	0.39	0.09	1.08	0.23	0.06	0.06	0.12	0.14
	(TS)	0.24	0.14	0.63	0.15	0.07	0.08	0.14	0.19
10000	**(R)**	0.25	0.04	0.72	0.15	0.03	0.03	0.07	0.06
	(TS)	0.15	0.07	0.41	0.09	0.04	0.05	0.08	0.11

Table 1: 100 × Squared Hellinger risk for proposed random penalty method and taut string

In many cases, the taut string or one of our proposals **(D)** and **(R)** is either best or the dyadic logarithm of relative risk w.r.t. the best method is close to zero. These three methods are 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 for the particular setting. The random

penalty **(R)** seems to be slightly better than the deterministic penalty **(D)** in many cases, the most notable exception being the trimodal unform density for $n = 50$. Cross-validation using the same set of partitions (i.e. a data-driven finest grid without further restrictions on minimum bin width) performs rather poorly, especially when the underlying density is a histogram (densities no. 1, 10, 13-16), has gaps in the support or regions where it is almost zero (5,10) or when it has infinite peaks (12). Note that it is particularly bad for the uniform, which can be a major problem in many applications like grey level estimation of image differences. Relative performance of **(CV)** w.r.t. the best method becomes generally worse when sample size increases. If we compare the random and deterministic penalties and cross-validation for the case of full dynamic programming optimization over a finest regular partition with bin length $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$ (**(Dr)**, **(Rr)**,**(CVr)**), the picture changes. Overall, the performance of all three methods is not bad, in particular **(CVr)** often outperforms **(Dr)** and **(Rr)**, which behave very similarly. An exception are again histogram densities, where cross-validation performs badly, especially for 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 the trimodal uniform density no. 10). The intermediate case, i.e using both penalties and cross-validation (**(Dc)**, **(Rc)**,**(CVc)**) for a data-driven finest grid but a adding the constraint that bin widths have to be at least $(x_{(n)} - x_{(1)}) \log^{1.5}(n)/n$ could be suspected to give a compromise between the finest regular grid suggested by theory and the greedy algorithm for a data-driven grid. However, **(Dc)** and **(Rc)** share the catastrophic behavior of **(Dr)** and **(Rr)** at the trimodal uniform density (no. 10) without offering a real improvement over **(D)** and **(R)** at the more well-behaved densities. On the other hand, **(CVc)** is a good compromise between **(CV)** and **(CVr)**, as it is in many cases either better than both or not far from the better of the two. It still shows bad behavior for the uniform and trimodal uniform densities. Table 3 shows that cross-validation has a pronounced tendency to choose histograms with a much larger number of bins than the penalized likelihood methods for all three sets of partitions. This is also illustrated by Fig. 2: For the uniform distribution with $n = 500$, our proposal **(R)** often chooses only one bin, which is the best possible for the uniform. **(CV)** chooses by far too many bins, resulting in a bad risk behavior. This is less extreme for **(CVr)**, but the number of bins chosen is still too large and the risk is more than 3 times larger than the best achieved for this setting.

Using AIC as a penalty leads to very bad results. It has already been shown theoretically in [6] and [26] and from a more practical point of view in [4] that AIC underpenalizes even for regular histograms. Since it does not account for the number of models of the same dimension, it is not surprising that this becomes even worse for the case of irregular histograms. Table 3 shows that the number of bins chosen on average is very often the largest among all methods considered. In many cases, the ratio of the Hellinger risk and the best risk achieved by any method in the simulation study is at least 4, often even much larger. BIC is a criterion which does not aim at a good control of risk but at asymptotically identifying the "smallest true model", if it exists. Although it also does not account for multiple models of the same dimension, it shows some good behavior in particular for small sample sizes that deteriorates when samples become larger. Particularly noteworthy is the

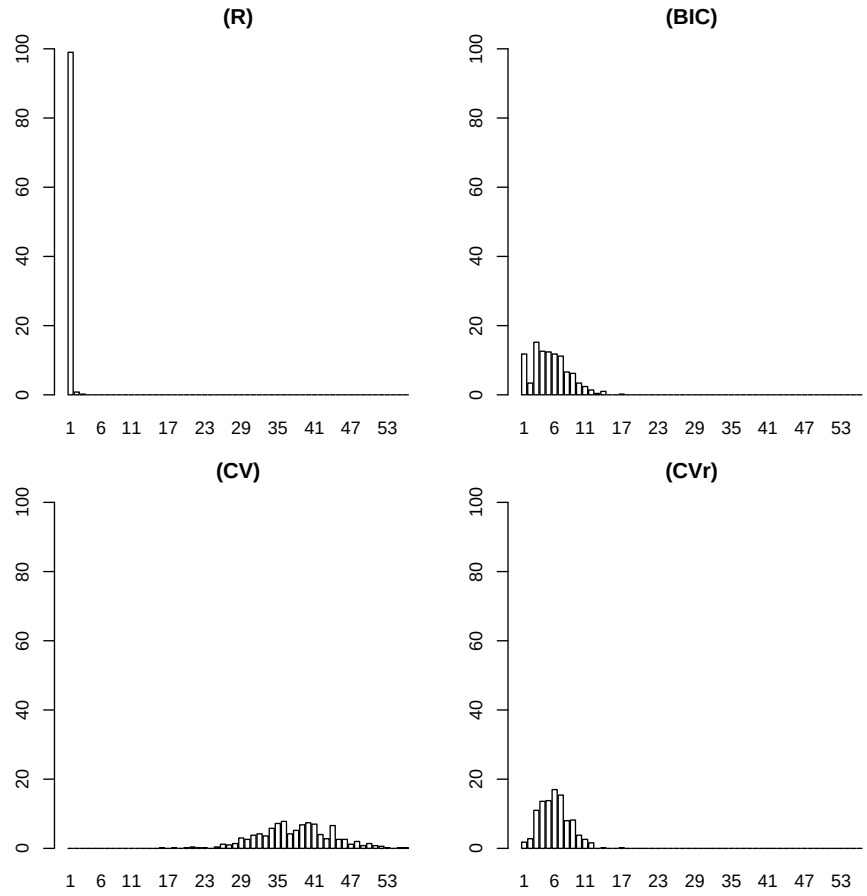

Figure 2: Barplots of number of bins chosen for the uniform density with $n = 500$ (percent of simulation runs).

bad performance for "simple" models like the uniform (which is also shown in Fig. 2) and the 5 bin regular histogram density no. 13.

The Birgé-Rozenholc construction of regular histograms **(BR)**, which improves on Akaike's penalization, compares quite favorably in many cases, being the best method for the normal, sawtooth and 5 bin regular histogram densities (nos. 4,11 and 13), at least when the sample size is not very small. This suggests that one does not always improve when choosing an irregular histogram instead of a regular one, since the greater flexibility may be outweighed by the greater difficulty in choosing a good partition, as was already remarked by Birgé and Rozenholc [4]. A regular histogram is of course inferior for spatially inhomogeneous densities like the lognormal (4) and the trimodal uniform (10).

The taut string method **(TS)** shows a particularly good behavior in terms of Hellinger risk. One should note here that 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. An example is given in Fig. 3, where the number of bins

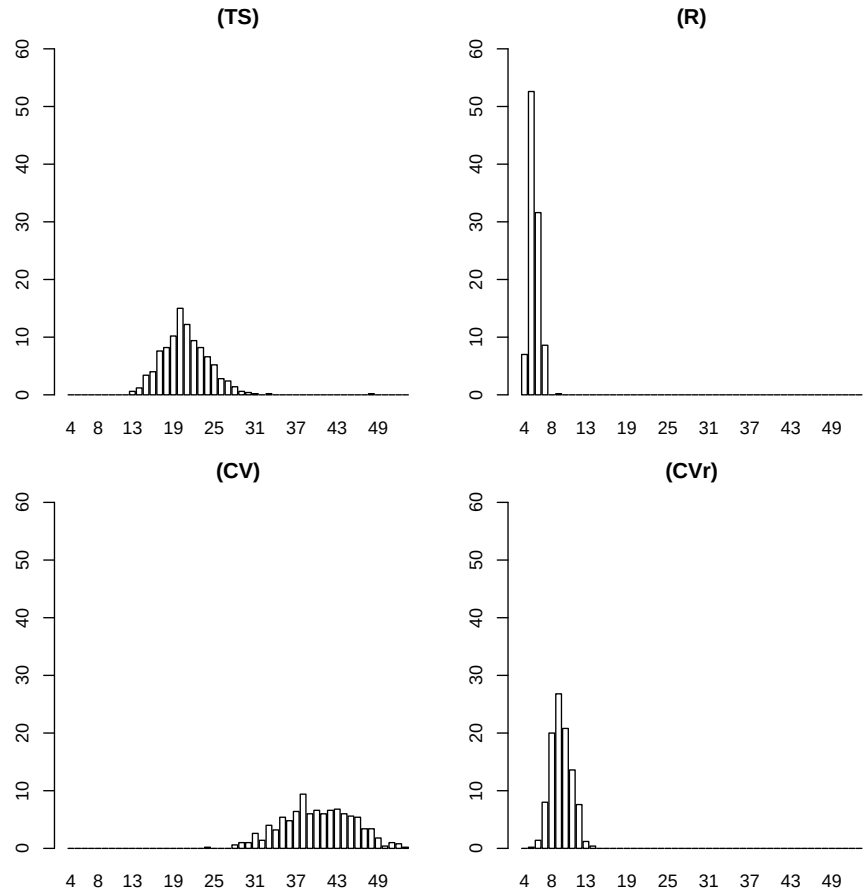

Figure 3: Barplots of number of bins chosen for the lognormal density $n = 500$ (percent of simulation runs).

chosen for the lognormal distribution (with $n = 500$) is shown. Of the four methods shown, **(CV)** performs worst, choosing again a large number of bins. **(TS)** is best in this scenario and uses a larger number of bins than both **(R)** and **(CVr)**.

Overall, between **(TS)** and **(R)** there is no clear winner, both often being the best method (in 36 and 16 of 96 cases, respectively, sometimes tied for the best). For some densities, **(R)** is better for small sample sizes while it is outperformed by **(TS)** for larger samples (e.g. densities 6, 7 and 11) while there are also cases like densities 1 and 16 where the opposite happens. One should note that the taut string method has originally been derived for different aims than achieving a good behavior w.r.t. a given loss function (see [13]), and many questions regarding behavior in a more classical framework remain open. It is also clear that the penalized likelihood approach can be more easily generalized to higher dimensions or not necessarily piecewise constant estimates, as has been done in a similar approach to nonparametric regression [11].

To summarize, we propose a practical method of irregular histogram construction in-

spired by theoretical works by Barron, Birgé and Massart [2], Castellan [6, 7] and Massart [26]. It can be easily implemented using a 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.

Acknowledgements

This work has been supported in part by the Collaborative Research Center Reduction of Complexity in Multivariate Data Structures (SFB 475) of the German Research Foundation (DFG). The authors also wish to thank Henrike Weinert for discussions and programming in earlier stages of the work.

A. Tables

den.	n	(D)	(R)	(CV)	(Dc)	(Rc)	(CVc)	(Dr)	(Rr)	(CVr)	(AIC)	(BIC)	(TS)	(BR)
f_1	50	0.13	0.05	1.53	0.03	0.03	0.76	0.00	0.00	0.44	2.07	1.27	0.07	0.12
	100	0.08	0.02	2.31	0.02	0.02	1.25	0.00	0.01	0.72	2.91	1.56	0.10	0.11
	500	0.06	0.03	4.20	0.01	0.02	2.56	0.00	0.01	1.69	4.64	2.04	0.19	0.13
	1000	0.04	0.02	4.64	0.01	0.01	3.05	0.00	0.01	2.30	4.98	2.10	0.12	0.11
	5000	0.01	0.00	5.14	0.00	0.00	3.68	0.00	0.00	3.83	5.37	1.89	0.12	0.12
	10000	0.00	0.00	5.38	0.00	0.00	3.97	0.00	0.00	4.71	5.59	1.85	0.20	0.18
f_2	50	0.42	0.34	0.50	0.38	0.34	0.09	0.52	0.47	0.00	0.91	0.53	0.13	0.14
	100	0.57	0.51	0.81	0.56	0.51	0.05	0.56	0.54	0.00	1.34	0.61	0.02	0.20
	500	0.71	0.67	1.79	0.70	0.66	0.04	0.69	0.67	0.07	2.20	0.75	0.00	0.20
	1000	0.74	0.70	1.94	0.72	0.70	0.08	0.72	0.70	0.15	2.25	0.73	0.00	0.18
	5000	0.73	0.72	1.78	0.73	0.71	0.23	0.71	0.70	0.47	1.94	0.62	0.00	0.11
	10000	0.69	0.71	1.59	0.69	0.71	0.21	0.67	0.68	0.68	1.70	0.56	0.00	0.07
f_3	50	0.51	0.42	0.68	0.49	0.43	0.14	0.51	0.48	0.00	1.15	0.64	0.51	0.23
	100	0.47	0.40	1.07	0.45	0.39	0.17	0.48	0.43	0.00	1.62	0.72	0.25	0.15
	500	0.65	0.60	2.08	0.64	0.60	0.36	0.65	0.59	0.10	2.49	0.81	0.13	0.00
	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	0.00
	5000	1.01	0.95	2.31	1.00	0.95	0.93	0.97	0.92	0.98	2.49	0.93	0.47	0.00
	10000	1.03	0.99	2.19	1.02	0.99	0.94	1.00	0.96	1.34	2.32	0.93	0.57	0.00
f_4	50	0.56	0.50	0.68	0.74	0.72	0.45	0.68	0.65	0.41	1.12	0.67	0.00	0.55
	100	0.63	0.60	1.08	0.72	0.72	0.38	0.70	0.69	0.38	1.62	0.79	0.00	0.61
	500	0.56	0.61	1.79	0.71	0.72	0.36	0.66	0.69	0.41	2.20	0.62	0.00	0.67
	1000	0.31	0.42	1.74	0.72	0.74	0.40	0.70	0.73	0.46	2.05	0.39	0.00	0.75
	5000	0.63	0.82	1.83	1.56	1.62	1.35	1.34	1.43	1.26	1.99	0.55	0.00	1.99
	10000	0.55	0.77	1.60	1.49	1.57	1.28	1.27	1.37	1.25	1.71	0.44	0.00	2.05
f_5	50	0.25	0.19	0.50	2.18	2.18	2.18	1.60	1.59	1.57	0.64	0.33	0.00	1.12
	100	0.47	0.44	0.76	2.71	2.71	2.73	1.81	1.82	1.77	1.07	0.48	0.00	0.41
	500	0.79	0.77	1.73	1.69	1.69	1.92	1.89	1.90	1.68	2.05	0.81	0.00	1.18
	1000	0.89	0.91	1.92	1.56	1.56	1.78	1.29	1.30	0.98	2.16	0.85	0.00	0.95
	5000	0.97	1.01	1.80	1.04	1.06	0.66	0.94	0.99	0.22	1.92	0.85	0.00	0.76
	10000	0.98	1.00	1.63	0.98	1.00	0.47	0.96	1.00	0.22	1.70	0.82	0.00	0.68
f_6	50	0.40	0.31	0.75	0.38	0.32	0.24	0.33	0.30	0.00	1.22	0.66	0.51	0.28
	100	0.36	0.29	1.14	0.35	0.30	0.31	0.45	0.39	0.00	1.70	0.80	0.44	0.31
	500	0.50	0.45	1.98	0.50	0.45	0.31	0.48	0.42	0.00	2.38	0.71	0.10	0.09
	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	0.00
	5000	0.88	0.82	2.24	0.88	0.82	0.84	0.83	0.78	0.92	2.42	0.84	0.00	0.15
	10000	0.92	0.87	2.09	0.91	0.87	0.84	0.91	0.86	1.26	2.23	0.83	0.00	0.11
f_7	50	0.28	0.16	0.48	0.25	0.17	0.04	0.39	0.35	0.00	0.86	0.45	0.39	0.19
	100	0.05	0.01	0.66	0.02	0.00	0.08	0.20	0.18	0.05	1.15	0.53	0.06	0.14
	500	0.68	0.64	1.44	0.66	0.63	0.14	0.69	0.66	0.15	1.84	0.68	0.00	0.27
	1000	0.89	0.87	1.80	0.88	0.86	0.44	1.15	1.14	0.29	2.10	0.80	0.00	0.45
	5000	1.16	1.13	1.93	1.15	1.13	0.68	1.13	1.10	0.82	2.07	1.07	0.00	0.74
	10000	1.25	1.24	1.84	1.25	1.24	0.76	1.26	1.26	1.11	1.93	1.11	0.00	0.77
f_8	50	0.30	0.24	0.08	0.29	0.26	0.05	0.34	0.32	0.00	0.40	0.21	0.22	0.14
	100	0.37	0.28	0.27	0.37	0.30	0.00	0.63	0.55	0.09	0.67	0.27	0.17	0.09
	500	0.44	0.36	0.90	0.51	0.46	0.24	0.66	0.59	0.00	1.25	0.35	0.03	0.10
	1000	0.47	0.39	1.05	0.54	0.49	0.27	0.64	0.58	0.09	1.33	0.37	0.00	0.19
	5000	0.81	0.75	1.18	0.87	0.82	0.52	0.86	0.80	0.41	1.31	0.69	0.00	0.46
	10000	0.99	0.93	1.16	1.01	0.96	0.61	0.99	0.93	0.66	1.25	0.82	0.00	0.70
f_9	50	0.44	0.42	0.00	0.45	0.45	0.16	0.44	0.44	0.36	0.34	0.03	0.26	0.36
	100	0.46	0.28	0.72	0.49	0.28	0.00	1.37	1.34	1.09	1.13	0.47	0.17	0.60
	500	0.55	0.50	1.81	0.52	0.48	0.36	1.64	1.62	1.36	2.17	0.68	0.00	0.85
	1000	0.61	0.56	2.01	0.59	0.55	0.50	1.65	1.62	1.37	2.30	0.72	0.00	1.08
	5000	0.74	0.70	2.01	0.73	0.70	0.76	1.66	1.64	1.58	2.16	0.72	0.00	1.23
	10000	0.73	0.69	1.86	0.72	0.69	0.74	1.61	1.59	1.69	1.98	0.65	0.00	1.31
f_{10}	50	0.00	1.31	0.55	3.15	3.15	3.14	3.12	3.12	3.12	0.68	0.31	0.15	2.68
	100	0.06	0.00	1.02	4.20	4.20	4.20	4.02	4.02	4.02	1.32	0.53	0.32	3.29
	500	0.02	0.00	2.40	5.82	5.82	5.82	5.64	5.64	5.64	2.73	0.79	0.51	5.08
	1000	0.01	0.00	2.79	6.80	6.80	6.80	6.34	6.34	6.34	3.06	0.81	0.56	5.63
	5000	0.00	0.00	3.29	9.09	9.09	9.09	7.57	7.57	7.57	3.45	0.73	0.73	0.17
	10000	0.00	0.00	3.43	10.10	10.10	10.10	6.46	6.46	6.46	3.54	0.63	0.81	0.17
f_{11}	50	0.06	0.02	0.54	0.01	0.01	0.22	0.00	0.00	0.13	0.81	0.51	0.19	0.03
	100	0.02	0.01	0.39	0.00	0.00	0.24	0.00	0.00	0.13	0.74	0.43	0.25	0.03
	500	0.86	0.61	0.80	0.89	0.70	0.27	1.06	1.06	0.15	1.14	0.48	0.11	0.00
	1000	0.63	0.58	0.98	0.62	0.59	0.40	0.86	0.69	0.08	1.24	0.63	0.09	0.00
	5000	0.97	0.90	0.99	0.96	0.90	0.45	0.98	0.90	0.24	1.08	0.84	0.13	0.00
	10000	1.06	0.99	0.92	1.05	0.98	0.57	1.09	1.01	0.37	0.97	0.91	0.17	0.00
f_{12}	50	0.21	0.11	0.88	0.01	0.02	0.25	0.00	0.01	0.14	1.36	0.79	0.03	0.15
	100	0.23	0.17	1.17	0.13	0.14	0.27	0.12	0.13	0.00	1.75	0.73	0.15	0.16
	500	0.37	0.33	2.41	0.30	0.30	0.88	0.29	0.28	0.28	2.83	0.90	0.00	0.29
	1000	0.52	0.49	2.63	0.48	0.47	1.09	0.52	0.51	0.50	2.95	0.89	0.00	0.36
	5000	0.65	0.62	2.61	0.64	0.62	1.24	0.63	0.62	1.22	2.80	0.80	0.00	0.52
	10000	0.68	0.66	2.48	0.67	0.66	1.18	0.67	0.65	1.61	2.62	0.74	0.00	0.57
f_{13}	50	0.10	0.07	0.69	0.04	0.05	0.02	0.01	0.02	0.00	1.16	0.65	0.06	0.06
	100	0.43	0.40	1.21	0.39	0.38	0.25	0.38	0.38	0.04	1.77	0.87	0.40	0.00
	500	1.41	1.34	3.43	1.39	1.33	1.86	1.53	1.50	1.41	3.85	1.88	1.10	0.00
	1000	1.06	0.98	3.98	1.02	0.98	2.44	0.83	0.78	1.70	4.29	1.97	1.11	0.00
	5000	0.91	0.90	4.66	0.90	0.90	3.32	1.20	1.20	3.35	4.87	1.96	1.13	0.00
	10000	0.92	0.91	4.82	0.90	0.91	3.56	1.39	1.39	4.10	4.99	1.92	1.25	0.00
f_{14}	50	0.18	0.13	0.65	0.12	0.11	0.14	0.12	0.13	0.00	1.10	0.59	0.19	0.02
	100	0.26	0.23	1.00	0.24	0.23	0.30	0.36	0.33	0.19	1.54	0.65	0.29	0.00
	500	0.08	0.03	2.41	0.02	0.00	0.85	0.58	0.56	0.67	2.83	0.83	0.09	0.88
	1000	0.14	0.11	2.98	0.12	0.11	1.45	0.83	0.82	1.25	3.30	0.98	0.00	0.94
	5000	0.01	0.00	3.76	0.00	0.00	2.36	1.41	1.42	2.68	3.96	1.07	0.38	1.90
	10000	0.01	0.01	3.95	0.00	0.00	2.66	0.91	0.92	3.30	4.12	1.03	0.56	2.63
f_{15}	50	0.10	0.00	0.06	0.35	0.29	0.06	0.46	0.42	0.06	0.43	0.09	0.11	0.03
	100	0.27	0.18	0.65	0.24	0.21	0.00	0.69	0.65	0.35	1.11	0.46	0.10	0.27
	500	0.15	0.00	1.79	0.15	0.01	0.31	1.15	1.09	0.87	2.17	0.48	0.21	0.48
	1000	0.06	0.01	2.39	0.03	0.00	0.87	1.26	1.25	1.32	2.69	0.67	0.34	0.87
	5000	0.01	0.00	2.82	0.00	0.00	1.57	1.34	1.34	2.17	3.01	0.64	0.16	0.99
	10000	0.01	0.00	2.79	0.01	0.00	1.69	1.49	1.48	2.58	2.94	0.51	0.09	0.85
f_{16}	50	0.11	0.00	0.29	0.96	0.93	0.74	0.74	0.71	0.48	0.65	0.29	0.00	0.34
	100	0.10	0.03	0.72	0.00	0.01	0.00	0.90	0.91	0.78	1.20	0.46	0.02	0.55
	500	0.07	0.02	1.80	0.58	0.57	0.79	1.53	1.52	1.32	2.19	0.50	0.00	1.21
	1000	0.00	0.00	2.18	0.27	0.29	0.84	1.21	1.22	1.26	2.49	0.55	0.03	1.24
	5000	0.06	0.00	2.64	0.04	0.00	1.36	1.12	1.10	1.92	2.83	0.48	0.45	1.57
	10000	0.04	0.00	2.96	0.04	0.00	1.83	1.17	1.15	2.58	3.11	0.58	0.80	1.65

Table 2: Dyadic logarithms of relative squared Hellinger Risks w.r.t best method

den.	n	D	R	CV	Dc	Rc	CVc	Dr	Rr	CVr	AIC	BIC	TS	BR
f_1	100	1 *95*	1 *99*	10 *15*	1 *99*	1 *99*	3 *26*	1 *100*	1 *99*	1 *28*	21 *10*	3 *18*	1 *99*	1 *89*
	1000	1 *98*	1 *99*	51 *07*	1 *100*	1 *100*	15 *13*	1 *100*	1 *100*	9 *13*	77 *07*	3 *16*	1 *99*	1 *89*
	10000	1 *100*	1 *100*	71 *07*	1 *100*	1 *100*	26 *09*	1 *100*	1 *100*	59 *06*	97 *16*	1 *35*	1 *98*	1 *84*
f_2	100	3 *54*	3 *51*	11 *16*	3 *59*	3 *52*	5 *53*	3 *59*	3 *53*	6 *41*	24 *10*	6 *20*	8 *22*	5 *21*
	1000	9 *47*	9 *49*	57 *07*	9 *49*	9 *49*	15 *25*	9 *46*	9 *47*	17 *23*	86 *08*	14 *15*	33 *10*	20 *11*
	10000	20 *34*	20 *32*	85 *14*	20 *35*	20 *32*	41 *15*	20 *32*	20 *34*	59 *08*	98 *29*	26 *16*	95 *08*	61 *07*
f_3	100	3 *74*	3 *81*	9 *15*	3 *78*	3 *81*	5 *48*	3 *71*	3 *73*	5 *39*	21 *10*	5 *19*	4 *24*	4 *33*
	1000	7 *56*	7 *60*	55 *08*	7 *59*	7 *60*	16 *17*	7 *55*	7 *56*	14 *18*	84 *08*	11 *17*	25 *12*	14 *14*
	10000	15 *39*	16 *35*	84 *14*	15 *40*	16 *35*	38 *14*	15 *39*	16 *33*	60 *07*	98 *27*	21 *17*	75 *08*	37 *06*
f_4	100	3 *66*	3 *78*	10 *16*	3 *89*	3 *89*	4 *66*	3 *82*	3 *82*	4 *45*	23 *10*	6 *21*	10 *20*	6 *18*
	1000	8 *51*	7 *41*	57 *08*	6 *47*	6 *53*	10 *25*	7 *45*	6 *52*	14 *18*	84 *08*	14 *16*	28 *11*	24 *07*
	10000	17 *37*	16 *36*	84 *12*	15 *39*	14 *43*	26 *13*	15 *43*	14 *43*	41 *08*	96 *20*	22 *16*	81 *08*	189 *02*
f_5	100	5 *36*	5 *38*	11 *15*	2 *92*	2 *93*	2 *80*	3 *73*	3 *76*	4 *46*	22 *11*	8 *19*	12 *16*	21 *19*
	1000	11 *38*	11 *34*	55 *09*	2 *92*	2 *93*	2 *80*	10 *40*	10 *42*	15 *30*	85 *09*	16 *19*	42 *12*	59 *05*
	10000	24 *27*	24 *31*	86 *12*	24 *29*	24 *32*	37 *17*	25 *28*	24 *34*	49 *14*	98 *24*	31 *16*	122 *07*	138 *04*
f_6	100	3 *46*	3 *52*	9 *17*	3 *45*	3 *48*	5 *33*	2 *56*	2 *53*	4 *38*	18 *10*	6 *16*	4 *24*	3 *23*
	1000	6 *52*	6 *51*	55 *09*	6 *55*	6 *51*	17 *18*	6 *54*	6 *54*	14 *17*	82 *07*	11 *16*	21 *10*	15 *13*
	10000	14 *32*	14 *36*	80 *12*	14 *33*	14 *36*	38 *12*	14 *34*	14 *38*	61 *07*	98 *29*	20 *16*	70 *09*	44 *07*
f_7	100	3 *79*	3 *87*	12 *18*	3 *86*	3 *87*	5 *34*	3 *77*	3 *79*	5 *34*	22 *11*	8 *14*	4 *25*	4 *28*
	1000	13 *19*	14 *18*	59 *10*	13 *20*	14 *17*	19 *19*	5 *36*	5 *32*	22 *16*	84 *09*	19 *15*	47 *08*	24 *09*
	10000	27 *19*	26 *19*	87 *11*	27 *19*	26 *20*	50 *13*	25 *20*	25 *19*	74 *08*	99 *32*	35 *16*	134 *07*	108 *04*
f_8	100	3 *25*	4 *25*	14 *18*	3 *34*	3 *29*	6 *48*	1 *45*	3 *40*	7 *34*	25 *09*	9 *17*	9 *10*	9 *27*
	1000	16 *21*	15 *22*	63 *09*	14 *34*	14 *29*	23 *20*	14 *23*	14 *27*	26 *16*	88 *09*	23 *16*	51 *08*	32 *11*
	10000	37 *20*	38 *20*	91 *15*	36 *21*	38 *18*	59 *14*	35 *22*	36 *21*	89 *06*	99 *36*	47 *14*	165 *05*	180 *03*
f_9	100	5 *64*	5 *75*	11 *18*	5 *69*	5 *77*	6 *52*	1 *37*	3 *34*	6 *32*	22 *11*	8 *18*	8 *17*	11 *18*
	1000	9 *52*	9 *56*	58 *09*	9 *57*	9 *57*	16 *18*	9 *35*	9 *37*	16 *18*	85 *09*	14 *16*	28 *14*	11 *28*
	10000	17 *41*	17 *38*	85 *12*	17 *43*	17 *38*	40 *12*	18 *35*	18 *33*	64 *07*	99 *29*	24 *16*	76 *08*	33 *12*
f_{10}	100	5 *86*	5 *100*	13 *16*	2 *99*	2 *100*	2 *100*	5 *100*	5 *100*	5 *84*	24 *11*	8 *18*	9 *20*	19 *59*
	1000	5 *98*	5 *100*	53 *07*	3 *100*	3 *100*	3 *61*	5 *99*	5 *100*	5 *47*	77 *07*	7 *17*	14 *14*	141 *86*
	10000	5 *100*	5 *100*	78 *07*	3 *100*	3 *100*	6 *20*	7 *100*	7 *100*	9 *23*	87 *10*	5 *37*	17 *12*	402 *100*
f_{11}	100	1 *87*	1 *94*	16 *16*	1 *99*	1 *99*	5 *32*	1 *100*	1 *100*	2 *27*	26 *10*	8 *12*	1 *47*	1 *86*
	1000	21 *29*	21 *32*	66 *11*	20 *36*	20 *33*	31 *20*	18 *17*	20 *25*	35 *17*	94 *12*	27 *15*	82 *07*	49 *08*
	10000	52 *13*	53 *16*	94 *19*	53 *14*	53 *15*	78 *16*	49 *13*	52 *15*	111 *07*	99 *39*	63 *13*	241 *06*	139 *07*
f_{12}	100	2 *38*	1 *44*	10 *17*	1 *63*	1 *59*	5 *28*	1 *78*	1 *77*	3 *44*	21 *11*	6 *18*	1 *70*	1 *39*
	1000	4 *37*	4 *36*	55 *07*	3 *39*	3 *37*	17 *14*	3 *50*	3 *46*	13 *15*	80 *07*	10 *16*	12 *19*	9 *14*
	10000	9 *46*	9 *47*	82 *11*	9 *48*	9 *47*	31 *11*	9 *46*	9 *46*	69 *07*	98 *23*	14 *16*	37 *11*	27 *05*
f_{13}	100	1 *60*	1 *52*	10 *16*	1 *63*	1 *53*	5 *50*	1 *77*	1 *71*	5 *33*	19 *10*	7 *18*	1 *63*	5 *50*
	1000	5 *85*	5 *93*	53 *08*	5 *91*	5 *94*	18 *17*	5 *89*	5 *92*	14 *15*	83 *09*	9 *16*	9 *15*	5 *100*
	10000	5 *97*	5 *99*	78 *10*	5 *100*	5 *99*	30 *11*	5 *100*	5 *100*	64 *06*	97 *21*	5 *21*	19 *12*	5 *100*
f_{14}	100	2 *55*	2 *62*	11 *17*	2 *69*	2 *71*	5 *43*	1 *56*	2 *52*	4 *29*	21 *10*	7 *19*	1 *43*	3 *70*
	1000	5 *57*	5 *63*	54 *07*	5 *60*	5 *63*	17 *14*	5 *53*	5 *57*	12 *13*	80 *06*	8 *15*	12 *16*	23 *56*
	10000	5 *98*	5 *99*	79 *10*	5 *100*	5 *100*	31 *10*	5 *100*	5 *99*	60 *06*	97 *21*	5 *22*	20 *11*	23 *85*
f_{15}	100	4 *61*	4 *66*	12 *16*	4 *73*	4 *75*	6 *40*	4 *51*	4 *54*	6 *37*	24 *10*	8 *17*	8 *15*	9 *18*
	1000	5 *85*	5 *93*	53 *08*	8 *80*	8 *83*	16 *15*	5 *89*	5 *92*	14 *15*	83 *09*	9 *16*	9 *15*	5 *100*
	10000	8 *54*	8 *49*	77 *09*	8 *54*	8 *49*	35 *11*	9 *35*	9 *34*	69 *07*	98 *21*	12 *17*	40 *09*	10 *96*
f_{16}	100	4 *77*	4 *82*	11 *17*	4 *90*	4 *84*	5 *48*	4 *87*	4 *85*	6 *36*	21 *10*	8 *18*	8 *18*	9 *35*
	1000	7 *58*	7 *49*	54 *09*	8 *80*	8 *83*	16 *15*	7 *41*	8 *40*	17 *15*	84 *08*	12 *16*	21 *11*	49 *19*
	10000	10 *78*	10 *86*	79 *11*	10 *80*	10 *86*	35 *09*	10 *65*	10 *70*	71 *06*	98 *24*	12 *20*	41 *08*	100 *98*

Table 3: Modes of number of bins chosen. The numbers in italics give the frequency of the mode in percent.

References

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

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

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

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

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

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

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

[8] Catoni, O., 2002. Data compression and adaptive histograms. In: Cucker F. and Rojas J.M. (Ed.), Foundations of Computational Mathematics, Proceedings of the Smalefest 2000, pages 35-60. World Scientific, 2002.

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

[10] 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.

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

[12] Davies, P. L., Gather, U., Nordman, D. J., and Weinert, H., 2008. A comparison of automatic histogram constructions. To appear in ESAIM: Probability and Statistics.

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

[14] Davies, P.L. and Kovac, A., 2008. ftnonpar: Features and strings for nonparametric regression. R package version 0.1-83. http://www.maths.bris.ac.uk/ maxak/ftnonpar.html

[15] Devroye, L. and Györfi, L., 1985. Nonparametric density estimation: the L_1 view. John Wiley, New York.

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

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

[18] Hartigan, J.A., 1996. Bayesian histograms. In: Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (Ed.), Bayesian Statistics 5, 211-222.

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

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

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

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

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

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

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

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

[27] Mildenberger, T., Weinert, H. and Tiemeyer, S., 2008. benchden: 28 benchmark densities from Berlinet/Devroye (1994). R package version 1.0.1.

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

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

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