ESAIM: PS June 2009, Vol. 13, p. 181–196
DOI: 10.1051/ps:2008005

ESAIM: Probability and Statistics
www.esaim-ps.org

A COMPARISON OF AUTOMATIC HISTOGRAM CONSTRUCTIONS*

Laurie Davies[1], Ursula Gather[2], Dan Nordman[3] and Henrike Weinert[4]

Abstract. Even for a well-trained statistician the construction of a histogram for a given real-valued data set is a difficult problem. It is even more difficult to construct a fully automatic procedure which specifies the number and widths of the bins in a satisfactory manner for a wide range of data sets. In this paper we compare several histogram construction procedures by means of a simulation study. The study includes plug-in methods, cross-validation, penalized maximum likelihood and the taut string procedure. Their performance on different test beds is measured by their ability to identify the peaks of an underlying density as well as by Hellinger distance.

Mathematics Subject Classification. 62G05, 62G07.

Received April 26, 2007. Revised February 13, 2008.

1. Introduction

Let $\mathbf{x}_n = \{x_{(1)}, \ldots, x_{(n)}\}$ be an ordered sample of real data points of size n. The goal is to find an automatic procedure which based on the data delivers a histogram with an appropriate number and widths of bins. This is known as the problem of histogram construction. We use the term "automatic procedure" (or simply "procedure") to refer to a fully prescribed algorithm for constructing a histogram that requires no further input by a user.

Most available histogram procedures yield so called regular histograms having equal length bins. Even for a regular histogram, the choice of the proper bin length has no generally accepted automatic solution. The number of proposed regular histogram procedures has become amazingly large. In contrast, irregular histogram constructions, which may adapt to local variability and result in histograms of varying bin width, are more rare. This may be due to a reliance on classical statistical decision theory to drive model selection. That is, a decision theory framework has produced many regular histogram procedures but has applied less easily in selecting highly parametrised irregular histograms. While some irregular histogram constructions have been proposed, these typically depend on tuning parameters without default values. Thus, they are not fully automatic procedures [3,11,20]) In addition, regular histogram procedures are often fast while many irregular

Keywords and phrases. Regular histogram, model selection, penalized likelihood, taut string.

* *This work has been supported by the Collaborative Research Center "Reduction of Complexity in Multivariate Data Structures" (SFB 475) of the German Research Foundation (DFG).*

[1] Department of Mathematics, University Duisburg-Essen; Department of Mathematics, Technical University Eindhoven, Germany.

[2] Department of Statistics, Technische Universität Dortmund, Germany; gather@statistik.uni.dortmund.de

[3] Department of Statistics, Iowa State University, USA.

[4] Department of Statistics, Technische Universität Dortmund, Germany.

histograms come with more extreme computational demands involving exhaustive search routines ([3,18–20,28]). One exception is the taut string method of [7] which produces an automatic generally irregular histogram at a computational cost of $O(n \log n)$. We include this method here as well as a method motivated by the taut string procedure which produces a regular histogram using appropriate Kuiper metrics.

The construction of histograms is an excellent problem for comparing the different paradigms of model choice. Our goal is to examine the performance of existing histogram procedures, each motivated by varying concepts of model quality, through simulation. We consider only histogram constructions available through full procedures. With the exception of the taut string, this treatment does exclude most irregular histogram methods mentioned above that depend on tuning parameters. However, it is fair to say that there are no guidelines available for using these in practice and we wish to avoid inventing implementations which may not accurately reflect these methods. Instead, we focus our investigation also on how existing histogram procedures perform in terms of a visual measure of model quality. This model metric assesses the extent to which a histogram construction matches the shape of a data generating density in terms of modes. The ability to identify the peaks in underlying densities is graphically an important property that has not received much consideration among histogram procedures. In addition we consider Hellinger distance as a measure of histogram quality.

2. Histogram procedures

Almost all histogram procedures involve optimality considerations based on statistical decision theory, where the performance of any data-based histogram procedure $\hat{f}(x) \equiv \hat{f}(x \mid \mathbf{x}_n)$ is quantified through its risk

$$R_n(f, \hat{f}, \ell) = \mathrm{E}_f\Big[\ell(f, \hat{f})\Big] \tag{2.1}$$

with respect to a given, nonnegative loss function ℓ and a density f, that is assumed to have generated the data $\mathbf{x}_n$. One usually seeks the histogram procedure $\hat{f}$ that minimizes (2.1), which is then deemed optimal.

The choice of a loss ℓ is important for judging histograms. There are many possibilities which include L_r-metrics

$$\ell(f,g) = \left(\int_{\mathbb{R}} |f(x) - g(x)|^r \mathrm{d}x\right)^{1/r}, \ 1 \le r < \infty; \quad \sup_x |f(x) - g(x)|, \ r = \infty,$$

the squared Hellinger distance

$$\ell^2(f,g) = \frac{1}{2}\int_{\mathbb{R}} \left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 \mathrm{d}x$$

and the Kullback-Leibler discrepancy

$$\ell(f,g) = \int_{\mathbb{R}} \log\left(\frac{f(y)}{g(y)}\right) f(y)\mathrm{d}y \in [0,\infty].$$

Although the choice of a loss function is to some extent arbitrary, reasons can be put forward for using one or the other. Birgé and Rozenholc [4] argue that Kullback-Leibler divergence is inappropriate because $\ell(f, \hat{f}) = \infty$ whenever a histogram $\hat{f}$ has an empty bin. However, there are histogram methods based on AIC, [1], or cross-validation rules [14], which are derived from risk minimization with this type of loss. The L_2-loss is popularly used because the asymptotic risk from (2.1) can then often be expanded and analysed (*cf.* [39] and references therein). Barron, Birgé and Massart [3] rely on squared Hellinger distance for determining histograms. Devroye and Györfi [9] give arguments in favour of the L_1-metric.

The construction of a histogram on the data range $[x_{(1)}, x_{(n)}]$ is essentially the same for all histogram procedures. A histogram is of the form

$$f_m(x) \equiv f_{m,\mathbf{t}_m}(x) = \frac{N_1}{n}\mathbb{I}\{t_0 \le x \le t_1\} + \sum_{j=2}^{m} \frac{N_j}{n}\mathbb{I}\{t_{j-1} < x \le t_j\}, \tag{2.2}$$

with the bin positions as a sequence of $m+1$ knots $\mathbf{t}_m = (t_0, t_1, \ldots, t_m) \in \mathbb{R}^{m+1}$, $t_j < t_{j+1}$. The number of bins of the histogram is $m \in \mathbb{N}$. The corresponding bin frequencies are

$$N_{j,m} = \begin{cases} \big|\{i : x_{(i)} \in [t_0, t_1]\}\big| & j = 1, \\ \big|\{i : x_{(i)} \in (t_{j-1}, t_j]\}\big| & j = 2, \ldots, m. \end{cases} \tag{2.3}$$

For regular histograms with equisized bins we get

$$\mathbf{t}_m = t_0 + \frac{(t_m - t_0)}{m}(0, 1, \ldots, m) \in \mathbb{R}^{m+1}, \quad t_0 < t_m \in \mathbb{R} \tag{2.4}$$

and a regular histogram $\hat{f}_m^{\mathrm{reg}}(x)$, $x \in [x_{(1)}, x_{(n)}]$, with m bins is determined by the knots

$$\hat{t}_j = x_{(1)} + \frac{j(x_{(n)} - x_{(1)})}{m}, \ j = 0, \ldots, m. \tag{2.5}$$

The bin frequencies are estimated by $\hat{N}_{j,m}$ as in (2.3) with t_j replaced by $\hat{t}_j$. The so called "anchor position" of the histogram (see [36]) is thus chosen here as $x_{(1)}$. Regular histogram procedures reduce to rules for determining an optimal number m^{opt} of bins that minimizes some type of risk in selecting a histogram from (2.5):

$$R_n(f, \hat{f}_{m^{\mathrm{opt}}}^{\mathrm{reg}}, \ell) = \inf_{m \in \mathbb{N}} R_n(f, \hat{f}_m^{\mathrm{reg}}, \ell).$$

Three broad categories of regular histogram constructions which differ by the methods used for determining m^{opt} in (2.5) are described in Sections 2.1–2.3. We present the taut string procedure and a Kuiper-metric based method in Section 2.4.

2.1. **Plug-in methods**

Assuming a sufficiently smooth underlying density f, the asymptotic risk (2.1) of a histogram $\hat{f}_m^{\mathrm{reg}}$ from (2.5) can often be expanded and minimized to obtain an asymptotically optimal bin number m^{opt}, which often depends on a constant $C(f) > 0$ determined by f. For example, a bin number $m^{\mathrm{opt}} = C(f)n^{1/3}$ is asymptotically optimal for minimizing the L_r-risk for $1 \le r < \infty$ as well as the squared Hellinger distance, while $m^{\mathrm{opt}} = C(f)[n/\log(n)]^{1/3}$ is asymptotically optimal with the L_∞-risk (*cf.* [12,32,39] for L_2; [9,16], for L_1; [20] for Hellinger distance; [34] for L_∞). The estimation of unknown quantities in $C(f)$ yields a plug-in estimate $\hat{m}$ of m^{opt}. Additionally, expressions for $C(f)$ and estimates of $\hat{m}$ are often simplified by assuming an underlying normal density f (*cf.* [32]). We will consider in greater detail a more sophisticated kernel method proposed in [39] for estimating $\hat{m}$. The WAND procedure is defined here using the two-stage bin width estimator $\tilde{h}_2$ with $M = 400$ given in [39] and implemented by *dpih* in the R-package `KernSmooth` ([40]).

2.2. **Cross-validation**

Cross-validation (CV) attempts to directly estimate the risk $R_n(f, \hat{f}_m^{\mathrm{reg}}, \ell)$ in approximating f by $\hat{f}_m^{\mathrm{reg}}$. This empirical risk can then be minimized by an estimate $\hat{m}$. In particular, the data $\mathbf{x}_n$ are repeatedly divided into two parts, one of which is used to fit $\hat{f}_m^{\mathrm{reg}}$ and the other to evaluate an empirical loss. These repeated loss evaluations can be averaged to estimate the risk $R_n(f, \hat{f}_m^{\mathrm{reg}}, \ell)$. Based on loss functions evoked by their names, L_2 cross-validation (L2CV) and Kullback-Leibler (KLCV) procedures require maximization of

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

respectively (*cf.* Rudemo [30], L2CV; Hall [14], KLCV).

2.3. **Penalized-maximum likelihood**

Many regular histogram procedures determine a histogram $\hat{f}^{\mathrm{reg}}_{\hat{m}}$ from (2.5) based on a bin number $\hat{m}$ that maximizes a penalized log-likelihood

$$\mathcal{L}_n(m) = \sum_{j=1}^{m} N_{j,m} \log(mN_{j,m}) - \mathrm{pen}(m). \tag{2.6}$$

Apart from an irrelevant constant, the first sum above corresponds to the log-likelihood $\sum_{j=1}^{n} \log(\hat{f}^{\mathrm{reg}}_m(x_{(j)}))$ of the observed data. The value $\mathcal{L}_n(m)$ is viewed as a single numerical index that weighs a regular histogram's fit to the data, as measured by the likelihood, against its complexity measured by the penalty term. The final histogram $\hat{f}^{\mathrm{reg}}_{\hat{m}}$ is judged to achieve the best balance between model fit and model complexity.

The penalty in (2.6) heavily influences the histogram $\hat{f}^{\mathrm{reg}}_{\hat{m}}$ and numerous choices have been proposed:

$$\mathrm{pen}(m) := \begin{cases} m & \text{AIC,} \\ m + \{\log(m)\}^{2.5} & \text{BR,} \\ m\log(n)/2 & \text{BIC,} \\ \log(C_{m,n}) & \text{NML.} \end{cases}$$

Akaike's Information Criterion (AIC), [1], is based on minimizing estimated Kullback-Leibler discrepancy. Birgé and Rozenholc [4] propose a modified AIC penalty (BR above) to improve the small-sample performance of the AIC procedure. The Bayes Information Criterion (BIC) follows from a Bayesian selection approach introduced in [31]. The Normalized Maximum Likelihood (NML) criterion uses an asymptotic ideal code length expansion (*cf.* [27]), derived by Szpankowski [38], as a penalty:

$$\begin{aligned} \log(C_{m,n}) \quad = \quad & \frac{m-1}{2}\log\left(\frac{n}{2}\right) + \log\left(\frac{\sqrt{\pi}}{(\Gamma(\frac{m}{2}))}\right) + \frac{\sqrt{2}m\Gamma(\frac{m}{2})}{3\sqrt{n}\Gamma(\frac{m-1}{2})} \\ & + \frac{1}{n}\left(\frac{3 + m(m-2)(2m+1)}{36} - \frac{m^2\Gamma^2(\frac{m}{2})}{9\Gamma^2(\frac{m-1}{2})}\right), \end{aligned}$$

where $\Gamma(\cdot)$ denotes the gamma function.

Rissanen [24–26] proposes several model selection techniques based on the principle of minimum description length (MDL). Information theory is applied to characterize the best model, with respect to a given model class, as the one providing the shortest encoding of the data $\mathbf{x}_n$. Hall and Hannan [15] apply different coding formulations to derive two further selection rules. To choose a bin number $\hat{m}$, the stochastic complexity (SC) and the minimum description length (MDL) procedures require maximization of

$$\frac{m^n(m-1)!}{(m+n-1)!}\prod_{j=1}^{m} N_{j,m}! \quad \text{or} \quad \sum_{j=1}^{m} N^*_{j,m}\log(N^*_{j,m}) - \left(n - \frac{m}{2}\right)\log\left(n - \frac{m}{2}\right) n\log(m) - \frac{m}{2}\log(n),$$

respectively, with $N^*_{j,m} = N_{j,m} - 1/2$.

2.4. **The taut string histogram procedure and a Kuiper-method**

The taut string procedure is described in [7], see also [6]. It assumes no true density and hence there is no loss or risk function. Instead it defines what is meant by an adequate approximation of the data and then attempts to find an adequate histogram with minimum number of peaks. This second step constitutes a kind of regularization.

We first give a brief description of the original taut string procedure. Let E_n denote the empirical distribution function of the data $\mathbf{x}_n$. Write the so called Kolmogorov tube of radius $\epsilon > 0$ centered at E_n as

$$T(E_n, \epsilon) = \Big\{ G;\ G : \mathbb{R} \to [0,1],\ \sup_x |E_n(x) - G(x)| \le \epsilon \Big\}.$$

The taut string function is best understood by imagining a string constrained to lie within the tube $T(E_n, \epsilon)$ and tied down at $(x_{(1)}, 0)$ and $(x_{(n)}, 1)$ which is then pulled until it is taut. There are several equivalent analytic ways of defining this. The taut string defines a spline function S_n on $[x_{(1)}, x_{(n)}]$ that is piecewise linear between knots $\{x_{(1)}, x_{(n)}\} \cup \{x_{(i)} : 1 < i < n, |S_n(x_{(i)}) - E_n(x_{(i)})| = \epsilon\}$, corresponding to points where S_n touches the upper or lower boundary of the tube $T(E_n, \epsilon)$. The knots define the bins and hence the histogram bin number, bin locations and bin probabilities as follows. If two consecutive knots are $x_{(i_j)}$ and $x_{(i_{j+1})}$, then the area of the bin $(x_{(i_j)}, x_{(i_{j+1})}]$ is proportional to the number of data points in $(x_{(i_j)}, x_{(i_{j+1})}]$ except for the first bin where the left point $x_{(1)}$ is included. The thus constructed taut string histogram s_n is known to have the fewest peaks or modes of any histogram whose integral lies in $T_n(E_n, \epsilon)$.

The size of the tube radius ϵ is important for the shape of the taut string histogram s_n. Davies and Kovac [7] prescribe a tube squeezing factor ϵ_n that determines the tube $T(E_n, \epsilon_n)$ and s_n as part of the taut string histogram procedure. This is done using a data approximation concept involving weak metrics applied to a continuous distribution function E and the empirical distribution E_n based on an i.i.d. sample from E. The κ-order Kuiper metric, $\kappa \in \mathbb{N}$, is defined by

$$d_{ku,\kappa}(E, E_n) = \sup \left\{ \sum_{j=1}^{\kappa} \big| \big(E(b_j) - E(a_j)\big) - \big(E_n(b_j) - E_n(a_j)\big) \big| : a_j \le b_j \in \mathbb{R},\ b_j \le a_{j+1} \right\}. \tag{2.7}$$

The definition of adequacy is based on the differences between successive Kuiper metrics $\rho_1(E, E_n) = d_{ku,1}(E, E_n)$ and $\rho_i(E, E_n) = d_{ku,i}(E, E_n) - d_{ku,i-1}(E, E_n)$ for $i > 1$. The distribution of $\rho_i(E, E_n)$, $i \in \mathbb{N}$, does not depend on E for continuous E. This can be seen as follows. If we denote a random i.i.d. sample from E by $X_1, \ldots, X_n$ then this can be generated as $X_i = E^{-1}(U_i)$ where E^{-1} denotes the inverse of E and the U_i are i.i.d. random variables uniformly distributed on $(0,1)$. It follows that $E(b_j) - E(a_j) = d_j - c_j$ and $E_n(b_j) - E_n(a_j) = \tilde{E}_n(d_j) - \tilde{E}_n(c_j)$ where $c_j = E(a_j)$, $d_j = E(b_j)$ and $\tilde{E}_n$ denotes the empirical distribution function of the U_i. As E is continuous $E((-\infty, \infty)) \supset (0,1)$ and hence the sup in (2.7) is taken over all c_j and d_j in $(0,1)$ with $c_j \le d_j \le c_{j+1}$. The above argument follows the proof of the corresponding result for the Kolmogorov metric. We now take $\kappa \ge 3$ to be a fixed odd integer. The software allows a maximum value of 19. Given this κ we choose $q_{\kappa,i}$, $i = 1, 2, \ldots \kappa$ and say that a taut string distribution S_n from a tube $T(E_n, \epsilon)$ provides an adequate data approximation if

$$\rho_i(S_n, E_n) \le q_{\kappa,i} \text{ for each } i = 1, \ldots, \kappa. \tag{2.8}$$

The $q_{\kappa,i}$ are chosen such that if E is the uniform distribution on $(0,1)$ then with probability 0.95 the resulting histogram has just the one bin. Asymptotically the chosen value of κ is irrelevant as the correct number of modal values will be found [7]. The finite sample performance does depend on κ. If it were known a priori that the density corresponding to E has k peaks then it is intuitively clear that $\kappa = 2k + 1$ would be the optimal choice. Failing this κ has to be fixed in advance. As it is clearly a hopeless task to find 10 peaks on a sample of size $n = 20$ from a ten-peaked distribution we choose κ as a function of the sample size n. In particular we put

$$\kappa = \kappa(n) = \begin{cases} 5 & n \le 50, \\ 9 & 51 \le n \le 100, \\ 19 & n \ge 101 \end{cases} \tag{2.9}$$

but clearly other choices are possible. In the taut string (TS) procedure we now reduce the tube radius ϵ of $T(E_n, \epsilon)$ until the approximation standard, $\rho_i(S_n, E_n) \le q_{\kappa,i}$, $i = 1, \ldots, \kappa$, is first met; this provides the squeezing factor ϵ_n to determine a final, usually irregular, taut string histogram s_n.

The software we use is available at http://www.stat-math.uni-essen.de/davies.html. The R-package `ftnonpar` [8] contains another version *pmden* of the TS algorithm which essentially has the same features.

Although not used in the remainder of the paper we mention that the version of the software applied here also includes a multiscale analysis of the data as developed by Dümbgen and Walther [10]. This removes one weakness of the Kuiper metric definition of adequacy which sometimes fails to pick up low power peaks centred on small intervals. If this is included then one bin only is returned in 90% (as against 95%) of the cases for uniformly distributed data.

Motivated by the taut string procedure we can derive a method which automatically produces regular histograms. We denote it by KUIP. This histogram results from computing histogram distribution functions for an increasing number of bins until all Kuiper-conditions, *i.e.* condition (2.8) above, are fulfilled.

3. Real data examples

We illustrate histogram construction with three data sets: eruptions of the Old Faithful geyser, the duration of treatment of patients in a suicide study, and the velocities of 82 galaxies from 6 well-separated conic sections of an unfilled survey of the Corona Borealis region. The first data set is found in [2], see also [13], the second in [35] and the last one in [23]. Extensive analyses by other authors have produced right skewed histograms for the suicide study data (cf. [33,35]) and histograms of varying modality for the galaxy data (*cf.* [23,29]). There are various versions of the Old Faithful data, which often produce histograms with two peaks; the version here (from geyser(MASS) in R) is heavily rounded with several observations identically "2" or "4".

Figures 1–3 provide histograms constructed with procedures from Section 2. The point made visually is the degree to which the histograms disagree in their shapes, largely when it comes to the number and position of peaks. We explore this aspect further in our numerical studies.

4. Simulation study

Our simulation study focuses on regular histogram procedures described in Section 2 as well as the irregular taut string procedure and the regular Kuiper-histogram. To limit the size of the study, we have excluded several regular histogram procedures involving plug-in estimates, such as Sturges's rule of $1+\log_2(n)$ bins [37], as well as methods from [5] and [17]. Numerical studies in [4] indicate that these are not competitive with the other methods that we consider.

We outline a new performance criterion in Section 4.1, motivated by the data examples in Figures 1–3. Section 4.2 describes the design of the simulation study to compare performances of histogram construction methods and the simulation results are summarized in Section 4.3.

4.1. Performance criterion: peak identification loss

We define a mode or peak of a density f as the midpoint of an interval $(x_1,x_2)\subset I\subset[x_1,x_2]$ which satisfies the following: $f(x)=c>0$ is constant on $x\in I$ and, for some $\delta>0$, it holds that $c>f(x)$ if $x\in I^\delta\setminus I$ for the enlargement $I^\delta=\cup_{y\in I}\{x\in\mathbb{R}:|x-y|\le\delta\}$ of I.

Identifying the locations of peaks in a reference density f is known to be a difficult problem for many histograms; see the discussion in [33] for the normal density. To illustrate this, Figure 4 provides histograms for a sample from the claw density, which is a normal mixture with five peaks taken from [22]. Two main errors in identifying peaks of the claw density f become evident in Figure 4. Histogram constructions can miss peaks of f (*e.g.*, BIC, MDL) or they can produce unnecessary peaks (*e.g.*, AIC). With these observations in mind, we propose the following loss to measure a histogram's performance in identifying peaks of a density f.

Suppose f is a density with $p=p(f)\in\mathbb{N}$ peaks at $z_1,\dots,z_p$ satisfying $(z_i-\delta_i,z_i+\delta_i)\cap(z_j-\delta_j,z_j+\delta_j)=\varnothing$, $i\ne j$, for some positive vector $\delta=\delta(f)\equiv(\delta_1,\dots,\delta_p)\in\mathbb{R}^p$. Assume that a histogram $\hat f$ has $\hat p=\hat p(\hat f)$ peaks at $y_1,\dots,y_{\hat p}$. We say a peak of $\hat f$ at y_j matches a peak of f at z_i if $\min_{1\le j'\le\hat p}|z_i-y_{j'}|=|z_i-y_j|<\delta_i$. An $\hat f$-peak

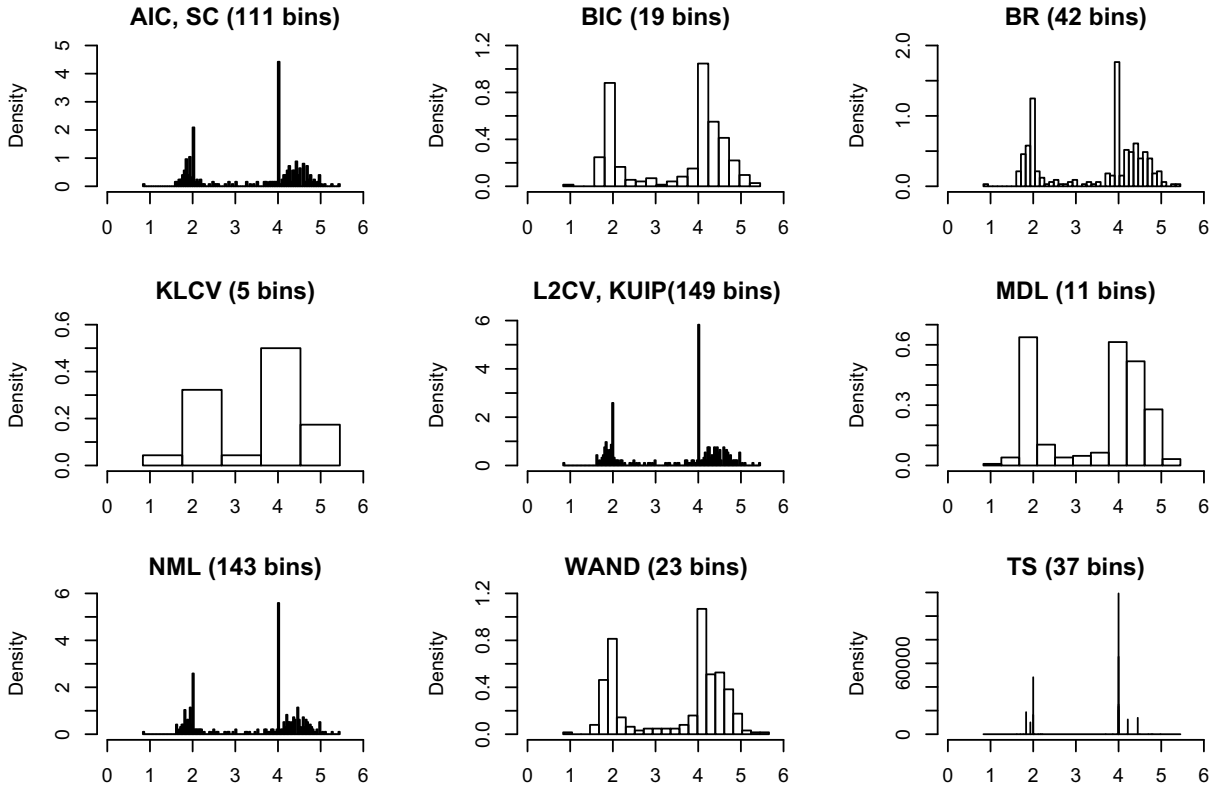

FIGURE 1. Histogram constructions from Old Faithful geyser data, $n = 299$.

that matches no peak of f is *spurious for* f while an f-peak that has no matches is said to be *unidentified by* $\hat{f}$. We can then define a peak identification loss as a count:

$$\begin{aligned} \ell_{\text{i.d.}}(f, \hat{f}, \delta) &= \#\text{ of unidentified peaks of } f \;+\; \#\text{ of spurious peaks of } \hat{f} \\ &= (p - C_{\text{i.d.}}) + (\hat{p} - C_{\text{i.d.}}) \end{aligned} \tag{4.1}$$

using the number $C_{\text{i.d.}} = \sum_{i=1}^{p} \mathbb{I}\{\min_{1\le j\le \hat{p}} |z_i - y_j| < \delta_i\}$ of correctly identified f-peaks. That is, the nonnegative loss $\ell_{\text{i.d.}}(f, \hat{f}, \delta) \ge 0$ measures the two possible errors incurred by identifying peaks of f with the peaks of $\hat{f}$. The vector δ represents the tolerances demanded in identifying each peak. Using $\ell_{\text{i.d.}}$ in (2.1), we obtain a risk for identifying peaks of a density f with a histogram procedure $\hat{f}$, which is a meaningful and interpretable measure of model quality.

4.2. **Simulation study design**

As test beds we select nineteen reference densities f of differing degrees of smoothness, tail behavior, support, and modality. The collection of reference densities includes: the standard Normal $N(0, 1)$, the Uniform $U(0, 1)$, standard Cauchy, triangle, and exponential distribution, eight mixture distributions from [22], a ten normal mixture used in [21], four densities, which are chosen to have roughly the same shapes as the test-case densities appearing in [4], and the "nearest unimodal density" to the bimodal density of [22].

The test densities are depicted in Figures 5 and 6. We use them for evaluating the performance of eleven histogram procedures: AIC, BIC, BR, KLCV, L2CV, MDL, NML, SC, WAND, KUIP, TS. We include the taut string histogram (TS) to consider this natural irregular competitor of KUIP. To measure the quality of

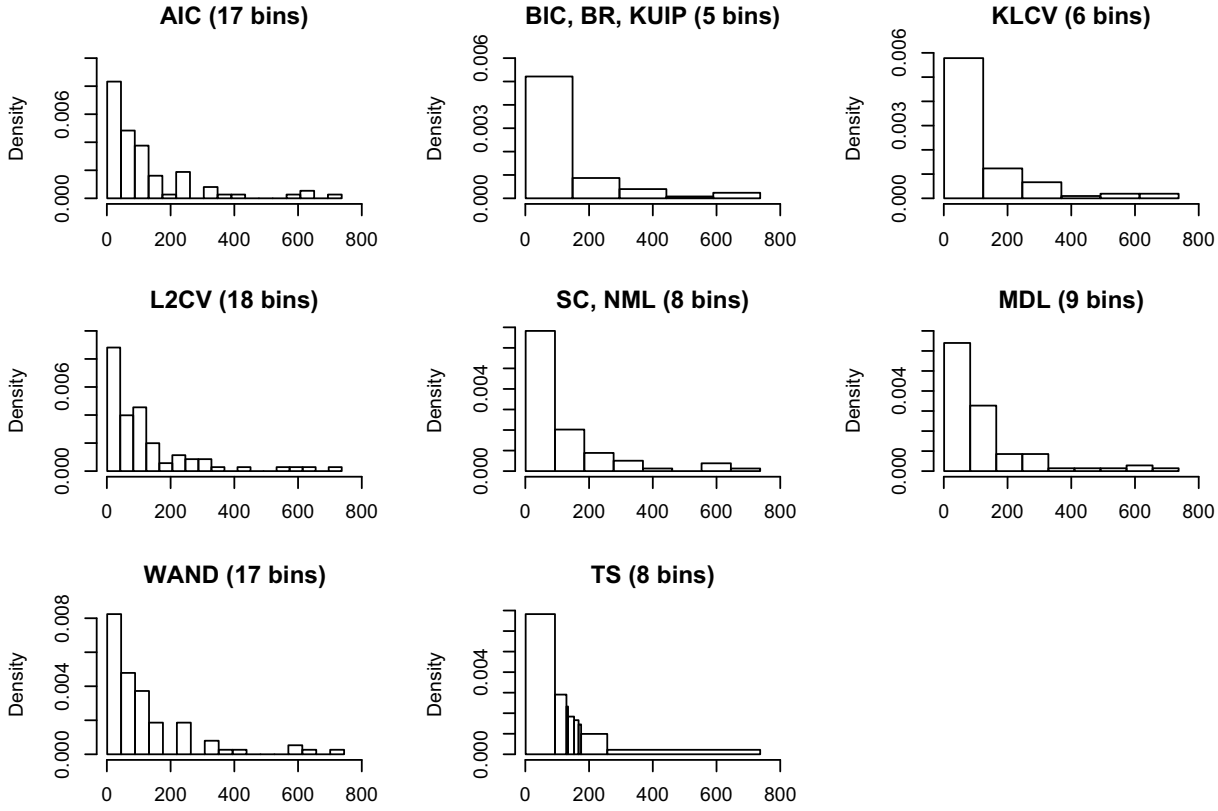

FIGURE 2. Histogram constructions from suicide data, $n = 86$.

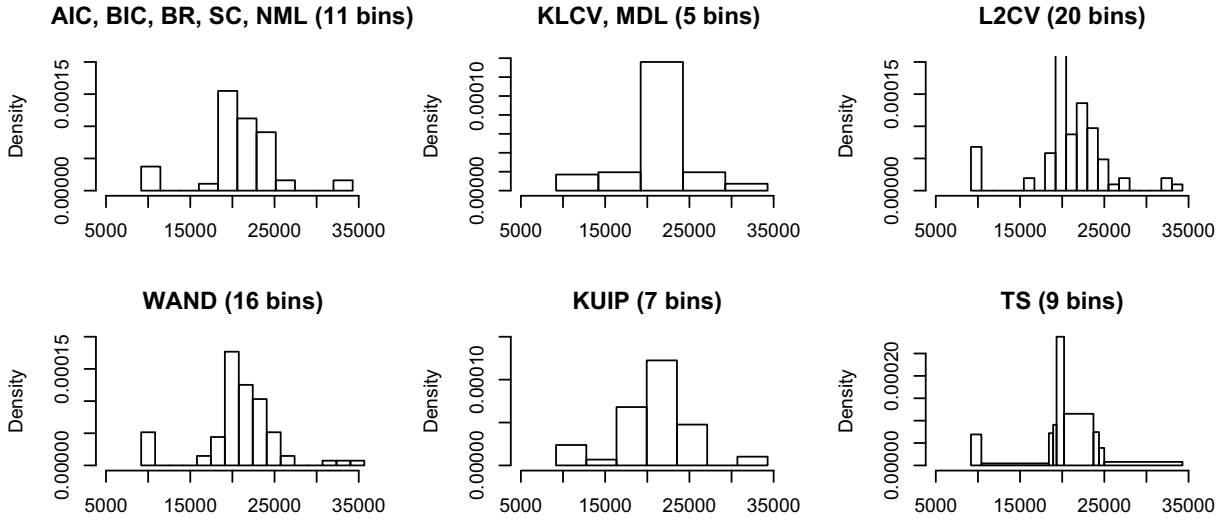

FIGURE 3. Histogram constructions from galaxy data, $n = 82$.

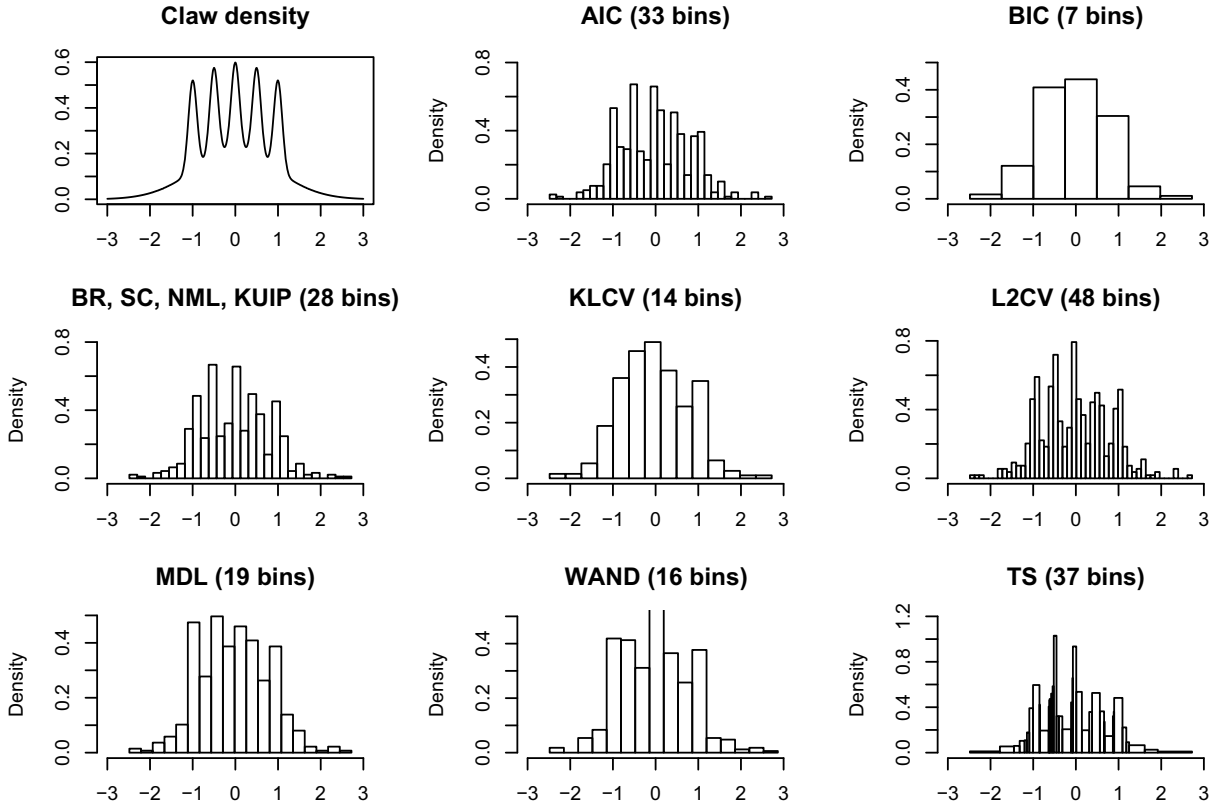

FIGURE 4. Histogram constructions for the claw density (first graph) based on a sample $n = 500$. The data sample was chosen here so that the histogram realisations have peak identification losses roughly matching the average losses (or risks) for each procedure in Table 1.

histograms, we consider risks based on two different losses: squared Hellinger distance and the peak identification (PID) loss from (4.1). The peak identification loss has an immediate interpretation, while the Hellinger loss seems appropriate for likelihood-based histograms. For each reference density f and sample size $n = (30, 50, 100, 500, 1000)$, we use 1000 independent size n samples $\mathbf{x}_{j,n} \equiv \mathbf{x}_{j,n}(f)$, $j = 1, \ldots, 1000$, to approximate the risk of each histogram procedure $\hat{f}$:

$$\widehat{R}_n(f, \hat{f}, \ell) = \frac{1}{1000} \sum_{j=1}^{1000} \ell(f, \hat{f}_{j,n}),$$

with loss evaluations $\ell(f, \hat{f}_{j,n})$ from histograms $\hat{f}_{j,n}$ at each simulation run $\mathbf{x}_{j,n}$.

4.3. **Simulation results**

Tables 1 and 2 provide the peak identification and Hellinger risks, respectively, for all procedures and for the sample sizes $n = 50, 500, 1000$.

To present a rough overview of the results of the simulation study, Figures 7–10 show the average ranks of the eleven histogram methods, resulting from ranking the procedures by their risks. In Figures 7 and 9 the ranks are calculated over all considered densities and all sample sizes of the simulation study; in Figures 8 and 10 the ranks are only taken over the unimodal and the multimodal densities separately.

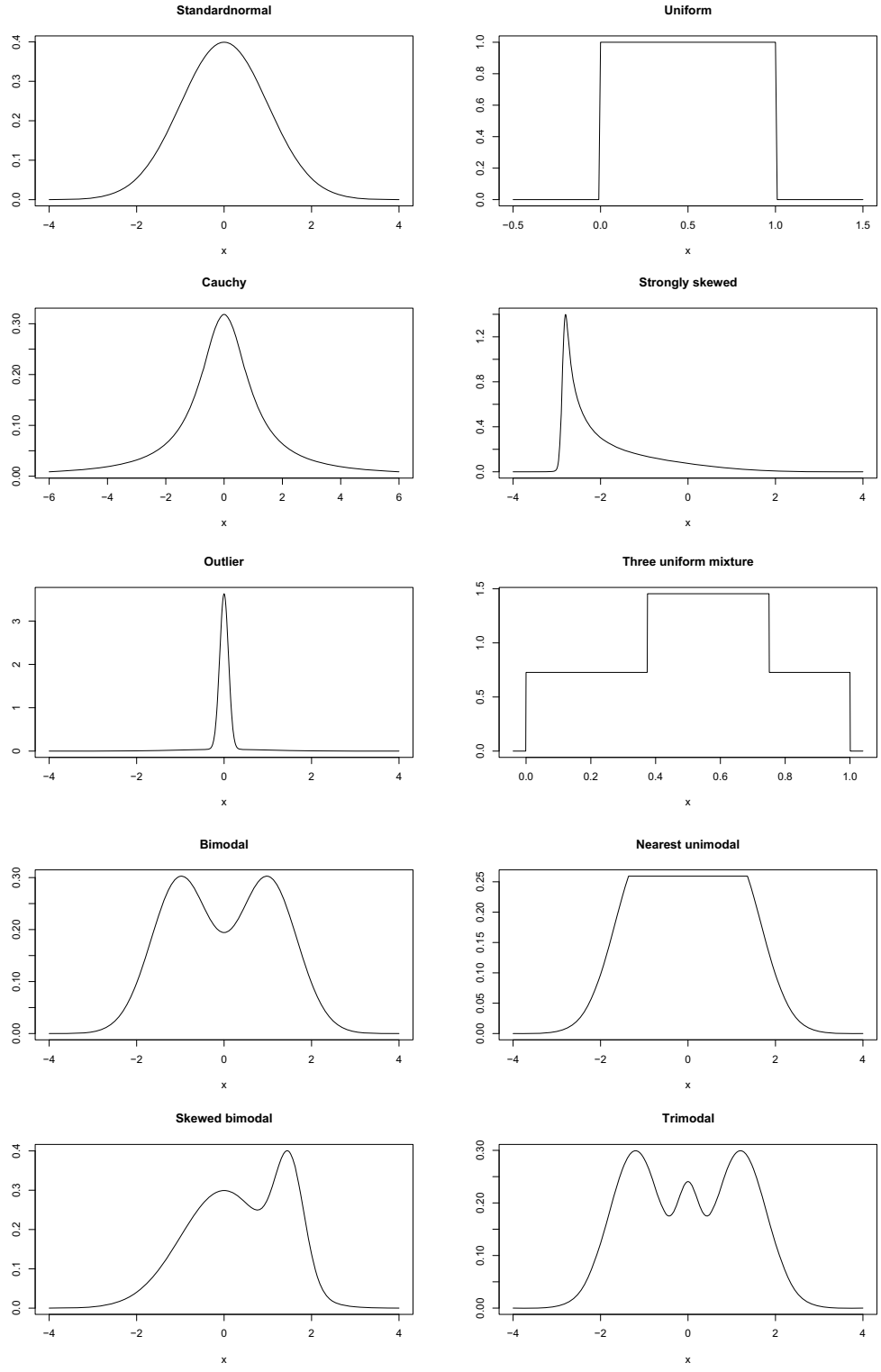

FIGURE 5. Data-generating densities used in the simulation study.

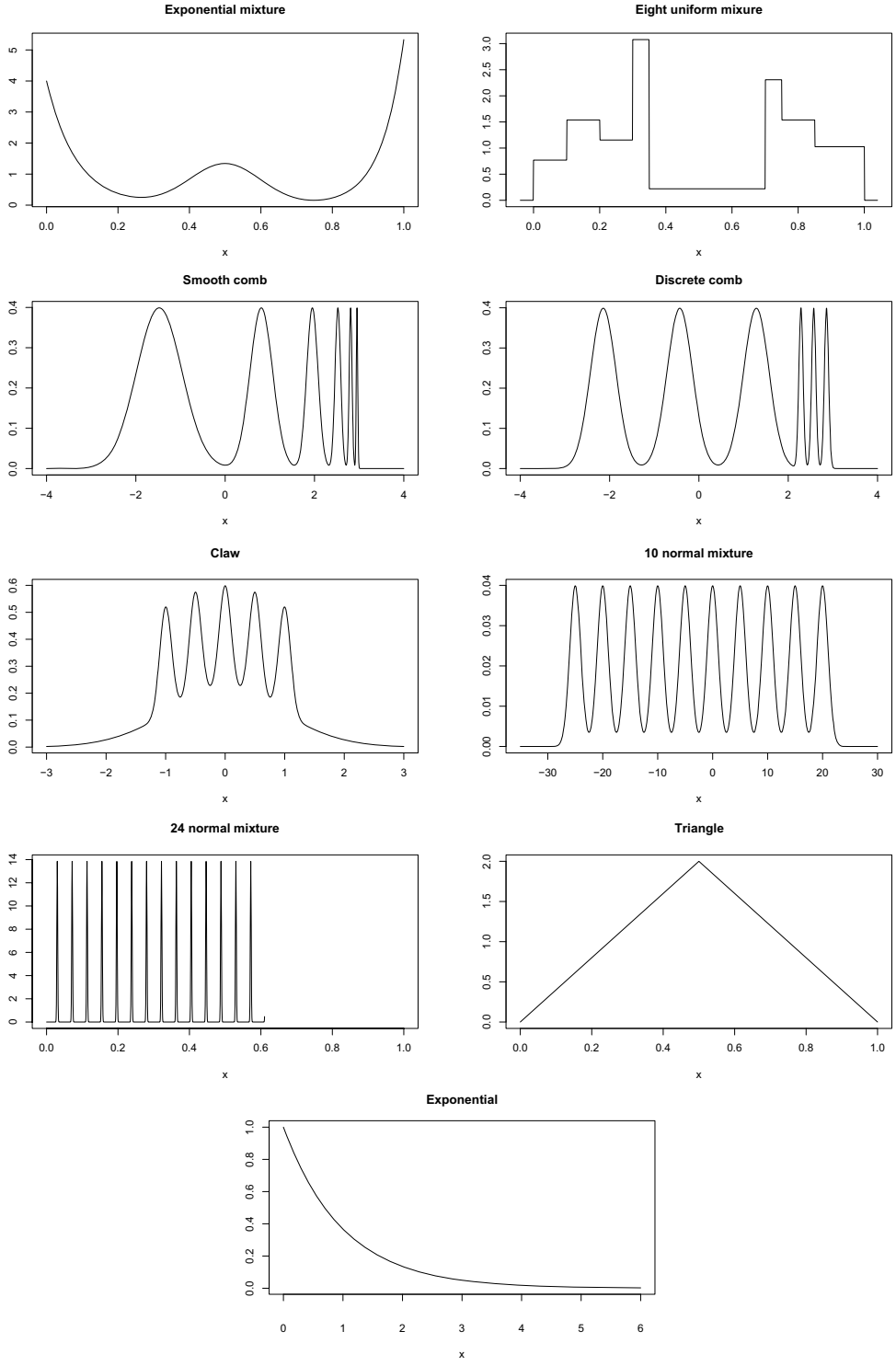

FIGURE 6. Data-generating densities used in the simulation study.

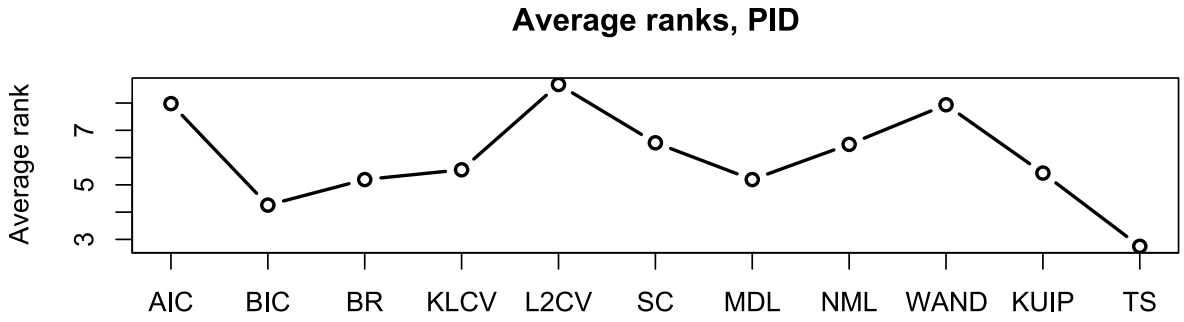

FIGURE 7. Average ranks for PID (all sample sizes and all densities).

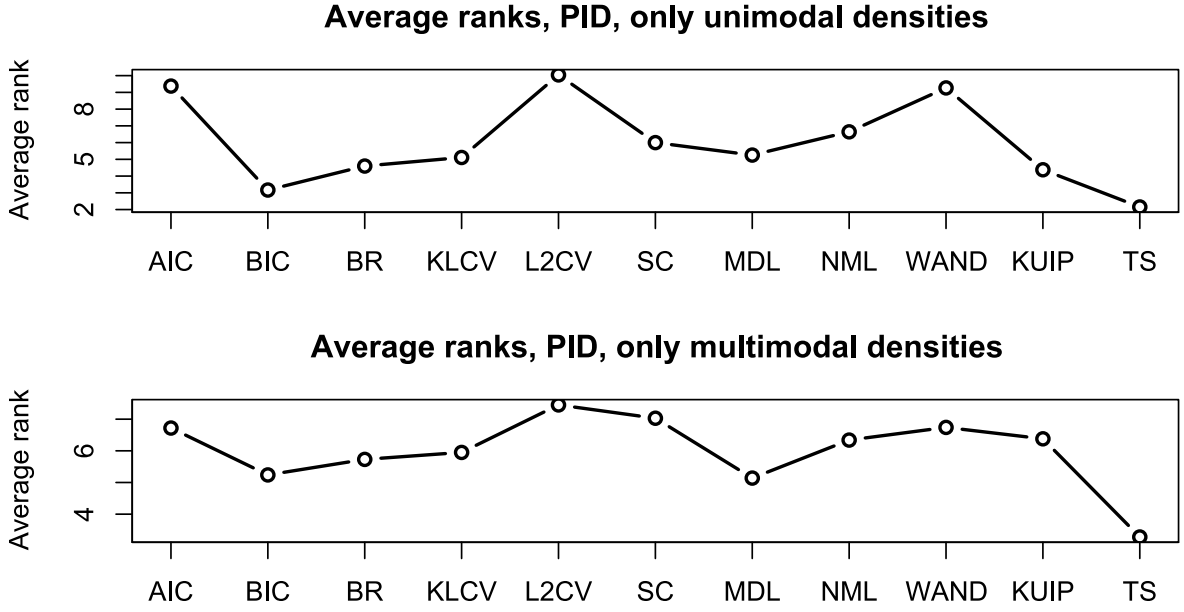

FIGURE 8. Average ranks for PID (all sample sizes) and unimodal or multimodal densities, respectively.

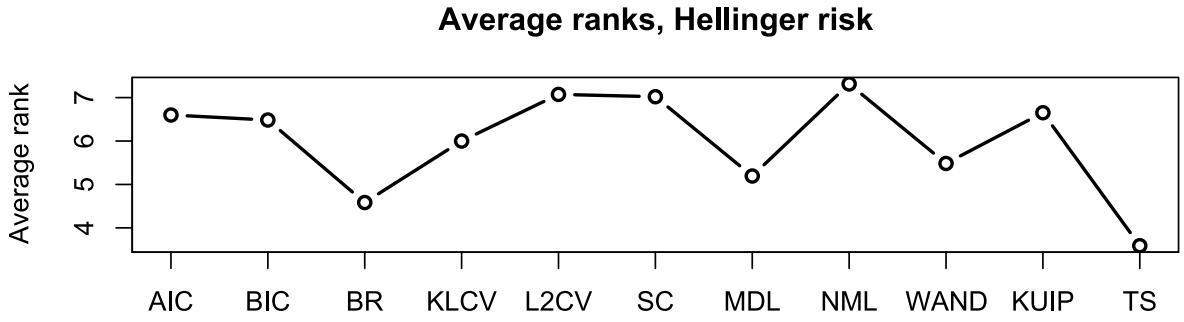

FIGURE 9. Average ranks for Hellinger loss (all sample sizes and all densities).

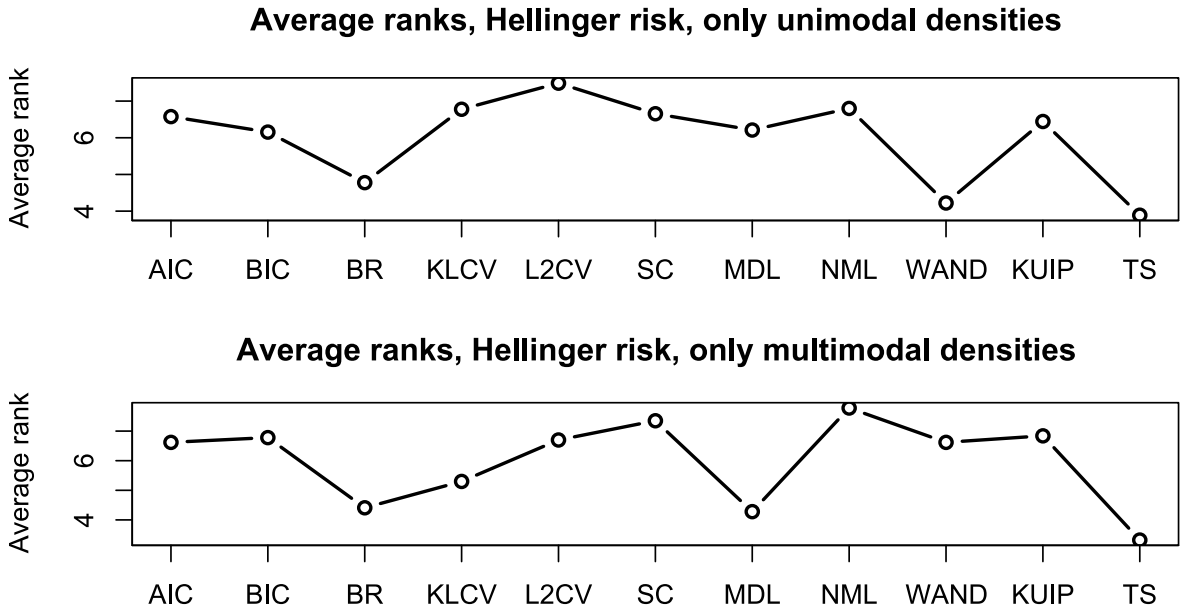

FIGURE 10. Average ranks for Hellinger loss (all sample sizes) and unimodal or multimodal densities, respectively.

From these figures, the differences among the average ranks for Hellinger loss are generally smaller than those for peak identification across the procedures. This indicates that more disagreement consistently emerges among the histograms when identifying modes rather than Hellinger distance. Such differences in peak identification are particularly evident in considering the ranks over unimodal densities (Fig. 8), where AIC, L2CV and WAND methods were generally the worst performers and BIC and TS methods performed relatively well. This demonstrates that regular histogram procedures generally tend to find too many peaks, which is also clear in examining the unimodal cases in Table 1. Such behavior can carry over to multimodal cases as well but some caution is required in interpreting the results. For illustration, we may consider the "bimodal" density in Table 1 where most regular histograms appear to perform well and BIC has the lowest peak identification risk of only 0.09 for samples of size $n = 500$. This compares to 1.37 for the taut string. However, if we look at the nearest unimodal density to the bimodal, shown in Figure 5, the performance of BIC deteriorates and its risk is now 0.82 as against 0.09 for the taut string; this pattern holds for many other regular histograms as well. The explanation is that the two peaks of the bimodal are not very pronounced and are difficult to find reliably. In this case, BIC seems to identify these peaks because it always tends to put two peaks there whether they are present in the density or not. The taut string cannot detect such weak peaks and has an error around 1. In fact no method can detect these peaks reliably, BIC only gives the illusion of doing so.

The irregular taut string histogram often exhibited the lowest Hellinger risk over many test densities and sample sizes, but differences were often relatively small compared to regular histograms (Tab. 2). The real weakness of regular histograms in being unable to adapt to local variability emerges most clearly in peak identification. The taut string procedure does not suffer from this weakness. This is also evident in comparing the taut string to its regular counterpart KUIP.

We can now summarize the results of the simulation study as follows:

- L2CV consistently is the worst performer in both peak identification and Hellinger distance.
- Of the information theory-based histograms, according to work of Rissanen [25,26], the NML and SC tend to perform similarly and are typically worse than MDL. Agreement between NML and SC also appears in the data examples (Figs. 2–4).

TABLE 1. Peak identification risk for histograms by density and sample size n.

density	n	AIC	BIC	BR	KLCV	L2CV	SC	MDL	NML	WAND	KUIP	TS
	50	0.68	0.06	0.06	0.10	0.85	0.93	0.50	0.57	0.18	0.06	0.01
$N(0,1)$	500	1.37	0.04	0.22	0.23	1.64	0.18	0.26	0.17	0.75	0.01	0.01
	1000	1.78	0.07	0.42	0.28	1.99	0.22	0.29	0.21	1.13	0.02	0.02
	50	0.76	0.04	0.06	0.31	0.56	0.23	0.38	0.35	1.00	0.05	0.05
$U(0,1)$	500	0.61	0.01	0.05	0.52	0.57	0.00	0.01	0.01	2.32	0.06	0.15
	1000	0.60	0.00	0.09	0.55	0.55	0.00	0.00	0.00	3.33	0.06	0.18
	50	2.67	2.30	2.25	1.68	3.75	2.47	1.57	2.70	4.58	2.66	0.00
Cauchy	500	8.64	6.32	6.87	1.98	12.64	7.50	2.06	8.69	19.59	14.05	0.00
	1000	12.45	8.93	9.59	2.02	18.73	10.82	2.11	12.53	13.86	20.57	0.01
	50	3.63	2.70	2.68	2.39	4.49	3.52	2.90	3.61	3.29	2.25	1.22
strongly skewed	500	8.42	2.42	4.05	2.08	13.92	3.81	2.74	3.91	8.60	3.89	0.09
	1000	12.41	3.17	7.86	2.19	17.38	5.82	3.51	6.13	12.31	5.67	0.01
	50	2.12	1.44	1.40	1.44	3.12	1.70	1.01	2.11	3.08	1.64	0.00
Outlier	500	10.70	4.83	9.01	1.28	17.45	6.59	1.77	9.14	17.66	10.92	0.01
	1000	15.43	8.45	13.30	1.24	26.01	10.13	2.20	12.29	26.44	13.99	0.01
	50	1.48	0.23	0.23	0.28	1.53	1.84	1.02	1.21	1.04	0.24	0.26
three uniform	500	0.64	0.00	0.00	0.46	0.94	0.00	0.01	0.00	3.54	0.10	0.07
	1000	0.58	0.00	0.00	0.52	0.93	0.00	0.00	0.00	4.99	0.09	0.07
	50	1.62	2.27	2.16	1.60	1.61	2.84	1.87	2.49	0.85	2.30	2.02
bimodal	500	1.02	0.09	0.15	0.32	1.07	0.16	0.18	0.11	0.25	0.20	1.37
	1000	1.43	0.02	0.27	0.44	1.37	0.16	0.20	0.13	0.52	0.03	0.48
	50	0.98	0.16	0.22	0.47	0.96	0.51	0.65	0.56	0.64	0.16	0.10
nearest	500	2.14	0.82	1.05	1.37	2.01	1.09	1.12	0.99	1.29	0.56	0.09
unimodal	1000	2.75	1.01	1.54	1.77	2.56	1.39	1.43	1.32	1.83	0.62	0.06
	50	1.33	1.05	1.06	1.00	1.34	1.59	1.10	1.49	1.10	1.11	1.04
skewed bimodal	500	1.88	0.46	0.60	0.65	1.95	0.56	0.58	0.55	0.97	0.74	1.00
	1000	2.51	0.33	0.87	0.59	2.65	0.55	0.58	0.57	1.66	0.39	0.92
	50	1.96	2.30	2.24	1.97	1.97	2.67	2.02	2.49	2.08	2.36	2.51
trimodal	500	1.26	0.93	0.80	0.77	1.31	0.79	0.78	0.81	0.63	1.09	1.53
	1000	1.34	0.61	0.49	0.62	1.24	0.45	0.44	0.46	0.46	0.77	1.06
	50	1.88	2.19	2.19	3.62	2.91	2.79	2.08	3.05	2.98	2.76	1.11
exp mixture	500	3.08	0.17	0.88	1.29	5.54	1.11	1.45	0.78	0.84	0.05	0.01
	1000	3.67	0.20	1.39	2.54	6.97	1.14	1.30	0.79	2.23	0.09	0.01
	50	4.81	4.31	4.29	4.11	4.96	5.41	4.38	5.18	4.45	4.14	4.00
eight uniform	500	4.13	3.54	3.65	3.58	4.67	3.71	3.75	3.64	3.95	4.44	1.72
	1000	3.72	3.38	3.41	3.45	3.91	3.41	3.41	3.40	5.34	3.05	0.94
	50	4.75	5.33	5.22	5.14	5.09	6.09	4.81	5.76	5.47	5.42	4.49
smooth comb	500	7.07	3.05	4.55	3.17	6.97	4.84	3.58	4.48	3.63	3.35	1.98
	1000	10.3	2.88	5.79	3.43	9.78	4.95	4.21	4.68	3.68	4.15	1.16
	50	4.03	4.57	4.39	4.91	4.21	4.96	3.82	4.78	5.73	4.68	3.83
discrete comb	500	4.36	3.73	2.74	3.56	4.75	2.88	2.88	2.80	3.38	3.06	0.98
	1000	7.24	2.15	2.51	2.13	7.12	2.01	1.62	1.76	2.46	2.35	0.29
	50	5.23	5.30	5.30	5.36	5.18	5.43	5.18	5.44	5.6	5.34	4.72
claw	500	5.47	5.35	4.12	4.98	7.10	4.14	4.41	4.25	5.07	3.99	0.34
	1000	7.44	4.08	4.46	4.65	10.17	3.51	3.72	3.72	4.11	4.36	0.02
	50	7.75	10.87	10.82	10.91	8.59	8.93	10.87	8.75	10.14	10.9	9.56
ten normal	500	3.18	3.99	2.31	2.47	2.72	2.76	2.18	2.41	10.45	2.50	0.57
	1000	2.83	1.8	1.42	1.24	2.32	1.51	1.36	1.18	8.58	1.07	0.22
	50	31.14	24.71	24.73	24.94	32.84	39.8	25.37	40.05	25.32	25.58	22.94
24 normal	500	1.44	10.31	34.45	30.67	2.10	1.41	32.12	1.27	27.14	17.22	0.24
	1000	0.52	0.92	4.03	31.28	1.70	0.23	34.18	0.40	28.07	2.15	0.02
	50	1.58	0.68	0.75	1.01	1.51	1.09	1.33	1.08	1.64	0.61	0.98
triangle	500	1.98	0.93	1.11	1.76	1.85	1.13	1.10	1.03	1.20	0.99	0.35
	1000	2.01	0.92	0.98	1.80	1.79	0.96	0.97	0.96	0.88	0.98	0.21
	50	1.45	0.55	0.54	0.67	2.41	1.04	0.62	1.14	1.60	0.53	0.01
exponential	500	2.98	0.90	1.45	0.61	5.50	1.15	0.78	1.35	7.78	0.68	0.02
	1000	3.82	1.17	2.16	0.64	7.38	1.40	0.87	1.70	12.04	1.19	0.03

- Among the regular histogram construction methods BIC, MDL and KUIP emerge (determined by average ranks in this order) as superior in terms of identifying the peaks of a density properly. BIC and KUIP show better results for unimodal densities while MDL has a small peak identification loss for multimodal densities (see Fig. 8).
- Among the regular histogram construction methods BR, MDL and WAND perform best in terms of Hellinger distance. BR yields small Hellinger risks consistently for unimodal and multimodal densities. WAND performs best for unimodal densities and MDL for multimodal densities (see Fig. 10).

TABLE 2. Hellinger risk*100 for histograms by density and sample size n.

density	n	AIC	BIC	BR	KLCV	L2CV	SC	MDL	NML	WAND	KUIP	TS
	50	5.13	4.68	4.59	4.31	5.33	6.84	4.60	5.60	3.41	5.44	4.96
$N(0,1)$	500	0.91	1.05	0.91	0.98	0.92	0.92	0.89	0.93	0.71	1.35	0.92
	1000	0.55	0.69	0.56	0.71	0.55	0.59	0.56	0.60	0.45	0.80	0.56
	50	3.76	2.18	2.27	2.57	3.18	2.60	2.77	2.92	4.90	2.09	2.29
$U(0,1)$	500	0.31	0.20	0.22	0.30	0.30	0.20	0.21	0.20	1.37	0.21	0.26
	1000	0.15	0.10	0.11	0.15	0.15	0.10	0.10	0.10	0.95	0.10	0.13
	50	12.78	14.96	17.15	39.22	12.22	12.61	29.67	12.15	10.60	13.49	5.84
Cauchy	500	9.02	14.03	15.47	65.61	7.99	9.99	57.25	8.81	3.28	9.23	1.17
	1000	7.87	13.87	15.29	72.38	6.56	9.07	65.21	7.65	1.09	7.89	0.73
	50	7.25	5.93	5.92	5.70	8.57	7.05	5.19	7.59	5.42	5.95	4.76
strongly skewed	500	2.04	2.36	1.99	2.93	2.30	2.00	2.32	2.01	1.93	2.29	0.99
	1000	1.32	1.80	1.29	3.06	1.41	1.39	2.05	1.38	1.19	1.75	0.59
	50	8.37	8.56	8.57	26.01	8.90	8.38	15.92	8.47	7.98	8.57	5.16
Outlier	500	2.88	3.75	2.93	15.87	3.44	3.29	7.76	2.93	3.39	2.85	1.02
	1000	1.92	2.47	1.92	13.22	2.45	2.12	6.20	1.95	2.46	1.83	0.69
	50	5.69	3.98	3.87	3.67	5.56	7.64	4.36	5.83	4.89	4.24	4.37
three uniform	500	0.45	0.36	0.36	0.43	0.48	0.36	0.36	0.36	1.35	0.38	0.49
	1000	0.23	0.18	0.19	0.22	0.24	0.18	0.18	0.18	0.98	0.19	0.27
	50	5.16	4.54	4.51	4.30	4.99	5.58	4.64	5.22	3.53	4.65	4.50
Bimodal	500	0.89	1.11	0.93	0.91	0.92	0.93	0.91	0.97	0.71	1.41	0.95
	1000	0.55	0.70	0.57	0.60	0.56	0.60	0.58	0.61	0.45	0.89	0.58
	50	5.00	4.37	4.37	4.18	5.01	5.34	4.50	5.18	3.40	4.49	4.39
nearest	500	0.86	1.05	0.91	0.87	0.93	0.91	0.90	0.94	0.70	1.49	0.81
unimodal	1000	0.52	0.68	0.55	0.56	0.55	0.57	0.56	0.59	0.44	0.99	0.49
	50	5.51	4.72	4.66	4.40	5.33	6.48	4.74	5.84	3.76	5.19	4.74
skewed bimodal	500	1.00	1.20	1.04	1.21	1.01	1.05	1.01	1.07	0.81	1.51	0.91
	1000	0.62	0.81	0.64	0.88	0.62	0.68	0.66	0.69	0.52	1.00	0.55
	50	5.26	4.32	4.33	4.30	5.08	5.33	4.57	5.21	3.71	4.34	4.32
trimodal	500	0.94	1.15	0.98	0.94	0.97	0.97	0.96	1.01	0.76	1.44	0.97
	1000	0.58	0.76	0.60	0.62	0.58	0.62	0.61	0.65	0.48	0.88	0.59
	50	9.43	8.08	7.89	7.35	11.01	11.11	6.05	11.49	12.96	7.00	6.58
exp mixture	500	1.36	1.50	1.30	1.20	1.46	1.28	1.26	1.31	3.22	1.57	1.00
	1000	0.77	0.94	0.76	0.73	0.83	0.78	0.77	0.80	2.23	0.99	0.59
	50	8.34	6.49	6.38	6.22	8.87	10.07	6.27	9.10	7.91	7.65	6.34
eight uniform	500	1.02	1.17	0.97	1.03	1.09	0.97	0.97	0.98	2.91	2.64	0.98
	1000	0.48	0.46	0.46	0.46	0.49	0.46	0.46	0.46	2.02	1.16	0.56
	50	9.97	10.14	10	9.63	10.39	12.7	8.67	12.32	10.55	10.32	9.40
smooth comb	500	2.58	3.08	2.51	3.12	2.57	2.46	2.60	2.52	3.36	2.92	2.25
	1000	1.78	2.16	1.70	2.24	1.75	1.72	1.76	1.74	2.35	2.07	1.34
	50	10.47	10.62	10.44	10.43	10.83	12.42	8.90	12.38	13.49	10.44	10.15
discrete comb	500	2.47	3.38	2.49	3.27	2.50	2.41	2.85	2.51	3.47	3.01	2.31
	1000	1.67	2.33	1.53	2.28	1.64	1.51	1.66	1.53	2.34	2.22	1.31
	50	7.59	6.47	6.39	6.21	8.27	9.17	6.06	8.58	5.46	7.26	6.28
claw	500	1.92	2.31	1.82	2.52	2.05	1.84	1.94	1.86	2.15	1.73	1.49
	1000	1.25	1.52	1.16	2.07	1.32	1.17	1.46	1.17	1.44	1.13	0.86
	50	12.86	8.19	8.25	8.55	12.10	13.56	8.74	14.14	10.34	8.39	8.91
ten normal	500	2.80	3.51	2.71	2.70	2.66	2.72	2.46	2.72	7.10	2.74	2.64
	1000	1.64	2.18	1.63	1.64	1.59	1.62	1.57	1.67	6.33	1.77	1.59
	50	59.81	62.70	62.72	62.83	58.28	53.68	62.86	53.51	63.68	63.10	60.63
24 normal	500	8.65	14.76	35.9	60.22	8.14	8.64	60.11	8.41	61.86	10.64	8.33
	1000	7.08	9.05	12.71	59.97	4.42	7.42	59.49	6.96	61.86	4.17	4.83
	50	4.83	4.25	4.15	3.86	4.64	5.70	4.46	5.13	3.20	4.32	4.50
triangle	500	0.77	0.89	0.79	0.75	0.79	0.79	0.80	0.82	0.60	1.03	0.82
	1000	0.46	0.59	0.49	0.45	0.48	0.50	0.51	0.53	0.38	0.65	0.53
	50	5.70	4.69	4.69	6.19	6.79	5.29	4.42	5.52	4.65	5.03	4.25
exponential	500	1.03	1.10	0.99	2.90	1.18	1.03	1.32	1.01	1.28	1.14	0.78
	1000	0.63	0.75	0.63	2.73	0.73	0.69	1.06	0.66	0.88	0.72	0.47

In particular, the WAND method in our simulations, based on two-stage estimation as described in [39], performed better in Hellinger distance than in simulations from Birgé and Rozenholc [4] (which used one-stage estimation described in [39]).

- The irregular TS method performs best in terms of Hellinger distance as well as in terms of peak identification for almost all densities.

As was to be expected, there is no overall optimal regular histogram procedure that delivers the best histogram for every data-generating density. However, the BR, BIC, WAND and MDL procedures provide good compromise methods for regular histogram construction over a wide range of densities and a variety of sample sizes. The BR method is better in terms of Hellinger risk and the BIC method with respect to detecting the peaks of a density. Altogether, the irregular TS histogram procedure is generally the best performer.

References

[1] H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automatic Control* **19** (1973) 716–723.
[2] A. Azzalini and A.W. Bowman, A look at some data on the Old Faithful geyser. *Appl. Statist.* **39** (1990) 357–365.
[3] A. Barron, L. Birgé and P. Massart, Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113** (1999) 301–413.
[4] L. Birgé and Y. Rozenholc, How many bins should be put in a regular histogram? *ESAIM: PS* **10** (2006) 24–45.
[5] J.E. Daly, Construction of optimal histograms. *Commun. Stat., Theory Methods* **17** (1988) 2921–2931.
[6] P.L. Davies and A. Kovac, Local extremes, runs, strings and multiresolution (with discussion). *Ann. Stat.* **29** (2001) 1–65.
[7] P.L. Davies and A. Kovac, Densities, spectral densities and modality. *Ann. Stat.* **32** (2004) 1093–1136.
[8] P.L. Davies and A. Kovac, `ftnonpar`, R-package, version 0.1-82, `http://www.r-project.org` (2008).
[9] L. Devroye and L. Györfi, *Nonparametric density estimation: the L_1 view.* John Wiley, New York (1985).
[10] L. Dümbgen and G. Walther, Multiscale inference about a density. *Ann. Stat.* **36** (2008) 1758–1785.
[11] J. Engel, The multiresolution histogram. *Metrika* **46** (1997) 41–57.
[12] D. Freedman and P. Diaconis, On the histogram as a density estimator: L_2 theory. *Z. Wahr. Verw. Geb.* **57** (1981) 453–476.
[13] I.J. Good and R.A. Gaskins, Density estimation and bump-hunting by the penalizes likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75** (1980) 42–73.
[14] P. Hall, Akaike's information criterion and Kullback-Leibler loss for histogram density estimation. *Probab. Theory Relat. Fields* **85** (1990) 449–467.
[15] P. Hall and E.J. Hannan, On stochastic complexity and nonparametric density estimation. *Biometrika* **75** (1988) 705–714.
[16] P. Hall and M.P. Wand, Minimizing L_1 distance in nonparametric density estimation. *J. Multivariate Anal.* **26** (1988) 59–88.
[17] K. He and G. Meeden, Selecting the number of bins in a histogram: A decision theoretic approach. *J. Stat. Plann. Inference* **61** (1997) 49–59.
[18] Y. Kanazawa, An optimal variable cell histogram. *Commun. Stat., Theory Methods* **17** (1988) 1401–1422.
[19] Y. Kanazawa, An optimal variable cell histogram based on the sample spacings. *Ann. Stat.* **20** (1992) 291–304.
[20] Y. Kanazawa, Hellinger distance and Akaike's information criterion for the histogram. *Statist. Probab. Lett.* **17** (1993) 293–298.
[21] C.R. Loader, Bandwidth selection: classical or plug-in? *Ann. Stat.* **27** (1999) 415–438.
[22] J.S. Marron and M.P. Wand, Exact mean integrated squared error. *Ann. Stat.* **20** (1992) 712–736.
[23] M. Postman, J.P. Huchra and M.J. Geller, Probes of large-scale structures in the Corona Borealis region. *Astrophys. J.* **92**, (1986) 1238–1247.
[24] J. Rissanen, A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **11** (1983) 416–431.
[25] J. Rissanen, Stochastic Complexity (with discussion). *J. R. Statist. Soc. B* **49** (1987) 223–239.
[26] J. Rissanen, *Stochastic complexity in statistical inquiry.* World Scientific, New Jersey (1989).
[27] J. Rissanen, Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42** (1996) 40–47.
[28] J. Rissanen, T.P. Speed and B. Yu, Density estimation by stochastic complexity. *IEEE Trans. Inf. Theory* **38** (1992) 315–323.
[29] K. Roeder, Density estimation with confidence sets exemplified by superclusters and voids in galaxies. *J. Amer. Statist. Assoc.* **85** (1990) 617–624.
[30] M. Rudemo, Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** (1982)65–78.
[31] G. Schwartz, Estimating the dimension of a model. *Ann. Stat.* **6** (1978) 461–464.
[32] D.W. Scott, On optimal and data-based histograms. *Biometrika* **66** (1979) 605–610.
[33] D.W. Scott, *Multivariate density estimation: theory, practice, and visualization.* Wiley, New York (1992).
[34] B.W. Silverman, Choosing the window width when estimating a density. *Biometrika* **65** (1978) 1–11.
[35] B.W. Silverman, *Density estimation for statistics and data analysis.* Chapman and Hall, London (1985).
[36] J.S. Simonoff and F. Udina, Measuring the stability of histogram appearance when the anchor position is changed. *Comput. Stat. Data Anal.* **23** (1997) 335–353.
[37] H. Sturges, The choice of a class-interval. *J. Amer. Statist. Assoc.* **21** (1926) 65–66.
[38] W. Szpankowski, On asymptotics of certain recurrences arising in universal coding. *Prob. Inf. Trans.* **34** (1998) 142–146.
[39] M.P. Wand, Data-based choice of histogram bin width. *American Statistician* **51** (1997) 59–64.
[40] M.P. Wand and B. Ripley, `KernSmooth`, R-package, version 2.22-21, `http://www.r-project.org` (2007).