Constructing a regular histogram - a comparison of methods

April 26, 2007

P. L. Davies[1], U. Gather[2], D. Nordman[3] and H. Weinert[2]

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 methods 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 the Hellinger distance and the ability to identify the modes of the underlying density.

Key Words: Regular histogram, model selection, penalized likelihood, taut-string

1 Introduction

Let $\mathbf{x}_n = \{x_{(1)}, \ldots, x_{(n)}\}$ be an ordered sample of real data points of size n. Using the data $\mathbf{x}_n$, the goal is to find a piecewise constant function that provides a "good simple approximation" of the data $\mathbf{x}_n$ on the sample range $[x_{(1)}, x_{(n)}]$ in the sense that integrating the function over each interval within $[x_{(1)}, x_{(n)}]$ approximates the relative frequency of data points within this interval with sufficient accuracy. This is known as the problem of histogram construction.

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

[2]Department of Statistics, University of Dortmund.

[3]Department of Statistics, Iowa State University.

We may classify existing procedures into two distinct categories, regular and adaptive histogram construction procedures. Regular histogram procedures produce only histograms with equal length intervals, while adaptive procedures may result in histograms with bins of varying lengths. The number of regular histogram procedures is amazingly large; in contrast, the number of available adaptive procedures is small. This may be due to a reliance on classical statistical decision theory to drive model selection. In particular, it seems to be very difficult both to choose the number and widths of the bins, if the latter are allowed to be different, in an automatic manner. Even if the aim is just to produce a regular histogram, the choice of the proper bin length has no generally accepted automatic solution.

The construction of regular histograms is an excellent problem for comparing the different paradigms of model choice. For regular histograms, the number of free parameters is the number of bins, one for this number m itself and $m-1$ for the bin occupancy numbers. In this paper we restrict attention to regular histograms, mainly because of the computational problems many methods have when the lengths of the bins are allowed to vary (Kanazawa (1988, 1993); Rissanen, Speed and Yu (1992); Barron, Birgé, and Massart (1999)). Furthermore, regular histogram methods are often attractive because they are typically fast and automatic, but in contrast not all adaptive procedures are automatic as some depend on tuning parameters without default values (Engel 1997, Kanazawa 1993 and Barron, Birgé and Massart 1999). One exception is the taut string method of Davies and Kovac (2004) which first produces an irregular histogram at a computational cost of $\mathrm{O}(n \log n)$. This can then be easily converted into a regular histogram by choosing a larger number of equal bins so that the shape of the histogram is not altered. For this reason, it competes with the other regular histogram methods on an equal footing, and we include it in the comparison.

2 Histogram procedures

Almost all regular 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

$$R_n(f, \hat{f}, \ell) = \mathrm{E}_f\left[\ell(f, \hat{f})\right] \tag{1}$$

with respect to a given, nonnegative loss function ℓ and an assumed density, denoted by f, that generates the data $\mathbf{x}_n$. One usually seeks the histogram procedure $\hat{f}$ that minimizes (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

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

squared Hellinger distance

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

and Kullback-Leibler discrepancy

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

The choice of a loss function is also quite arbitrary. For judging histogram quality, Birgé and Rozenholc (2006) 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, Akaike (1973), or cross-validation rules by Hall (1990), which are derived from risk minimization with this type of loss. The L_2 loss is popularly used because the asymptotic risk from (1) can then often be expanded and analyzed [c.f. Wand (1997) and references therein]. Barron, Birgé and Massart (1999) and Birgé and Rozenholc (2006) rely on squared Hellinger loss for determining histograms.
Construction of a histogram on the data range $[x_{(1)}, x_{(n)}]$ is essentially the same for all regular histogram procedures. A general histogram is of the form

$$f_m(x) \equiv f_{m,\mathbf{p}_m,\mathbf{t}_m}(x) = \frac{p_1}{t_1 - t_0}\mathbb{I}\{t_0 \le x \le t_1\} + \sum_{j=2}^{m} \frac{p_j}{t_j - t_{j-1}}\mathbb{I}\{t_{j-1} < x \le t_j\}, \quad (2)$$

with the histogram parameters

- the number $m \in \mathbb{N}$ of bins in the histogram,
- 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 corresponding bin probabilities $\mathbf{p}_m = (p_1, \ldots, p_m) \in [0,1]^m$, $\sum_{j=1}^m p_j = 1$.

The restriction to histograms with equal bins yields:

$$\mathbf{t}_m = t_0 + \frac{(t_m - t_0)}{m}(0, 1, \ldots, m) \in \mathbb{R}^{m+1}, \quad t_0 < t_m \in \mathbb{R}, \quad (3)$$

and a regular histogram $\hat{f}_m^{\text{reg}}(x)$, $x \in [x_{(1)}, x_{(n)}]$, with m bins is determined by the following knots $\hat{\mathbf{t}}_m$ and probabilities $\hat{\mathbf{p}}_m$:

$$\hat{t}_j = x_{(1)} + \frac{j(x_{(n)} - x_{(1)})}{m}, \quad \hat{p}_j = \frac{N_{j,m}}{n}, \quad N_{j,m} = \begin{cases} \left|\{i : x_{(i)} \in [\hat{t}_0, \hat{t}_1]\}\right| & j = 1, \\ \left|\{i : x_{(i)} \in (\hat{t}_{j-1}, \hat{t}_j]\}\right| & j = 2, \ldots, m. \end{cases} \quad (4)$$

The bin probabilities are usually estimated by the bin relative frequencies $\hat{p}_j$, corresponding to the maximum likelihood estimates given m and $\hat{\mathbf{t}}_m$.

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 (4):

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

With the exception of the taut-string, there are three broad categories of regular histogram procedures which differ by the methods used for determining m^{opt} in (4). We summarize these in Sections 2.1 - 2.3 and describe the taut-string method in Section 2.4.

2.1 Plug-in methods

Assuming a sufficiently smooth underlying density f, the asymptotic risk (1) of a histogram $\hat{f}^{\text{reg}}_m$ from (4) can often be expanded and minimized to obtain an asymptotically optimal bin number m^{opt}. For example, a bin number $m^{\text{opt}} = C(f)n^{1/3}$ is asymptotically optimal for minimizing the L_r risk for $1 \le r < \infty$ as well as the squared Hellinger distance, while $m^{\text{opt}} = C(f)[n/\log(n)]^{1/3}$ is asymptotically optimal with the L_∞ risk [c.f. Scott (1979), Freedman and Diaconis (1981), Wand (1997) for L_2; Devroye and Györfi (1985), Hall and Wand (1988) for L_1; Kanazawa (1993) for Hellinger distance; Silverman (1978) 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 [c.f. Scott (1979)]. We will consider in greater detail a more sophisticated kernel method proposed by Wand (1997) for estimating $\hat{m}$. As in Birgé and Rozenholc (2006), the WAND procedure is defined here using the one-stage bin width estimator $\tilde{h}_1$ with $M = 400$ given in formula (4.1) of Wand (1997).

2.2 Cross-validation

Cross-validation (CV) attempts to directly estimate the risk $R_n(f, \hat{f}^{\text{reg}}_m, \ell)$ in approximating f by $\hat{f}^{\text{reg}}_m$. 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 the model $\hat{f}^{\text{reg}}_m$ and the other to evaluate an empirical loss (e.g., delete-1 CV). These repeated loss evaluations can be averaged to estimate the risk $R_n(f, \hat{f}^{\text{reg}}_m, \ell)$. 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 [c.f. Rudemo (1982), L2CV; Hall (1990), KLCV].

2.3 Penalized-maximum likelihood

Many regular histogram procedures determine a histogram $\hat{f}^{\text{reg}}_{\hat{m}}$ from (4) based on a bin number $\hat{m}$ that maximizes a penalized log-likelihood:

$$\mathcal{L}_n(m) = \sum_{j=1}^{m} N_{j,m} \log(mN_{j,m}) - \text{pen}(m). \tag{5}$$

Apart from an irrelevant constant, the first sum above corresponds to the log-likelihood of the observed data $\sum_{j=1}^{n} \log(\hat{f}^{\text{reg}}_m(x_{(j)}))$. 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 $\hat{f}^{\text{reg}}_{\hat{m}}$ is judged to achieve the best balance between model fit and model complexity.
The penalty in (5) heavily influences the histogram $\hat{f}^{\text{reg}}_{\hat{m}}$ and numerous choices have been proposed:

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

Akaike's Information Criterion (AIC), Akaike (1973), is based on minimizing estimated Kullback-Leibler divergence. Birgé and Rozenholc (2006) 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 by Schwartz (1978). The Normalized Maximum Likelihood (NML) criterion uses an asymptotic ideal code length expansion [c.f. Rissanen (1996)], derived by Szpankowski (1998), as a penalty:

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

where $\Gamma(\cdot)$ denotes the gamma function. We discuss NML further along with similar coding-based selection rules below.
Rissanen (1983, 1987, 1989) 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 (1988) apply different coding formulations to derive two further selection rules for regular histograms $\hat{f}^{\text{reg}}_{\hat{m}}$. To choose a bin number $\hat{m}$, the stochastic complexity (SC) and minimum description length (MDL) procedures require maximization of

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

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

2.4 Taut-string (TS) histogram procedure

The taut-string method is presented in Davies and Kovac (2004). 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 the minimum number of peaks. This second step constitutes a kind of regularization. As mentioned above the taut-string histogram is not regular but it can yield a regular histogram by choosing a sufficiently large number of bins so that the regular and non-regular histograms differ only slightly.
We now give a brief description of the taut-string method.
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 $T(E_n, \epsilon)$ is best understood by imagining a string constrained to lie within the tube 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 $\{x_{(1)}\} \cup \{x_{(i)} : 1 < i < n,\ |S_n(x_{(i)}) - E_n(x_{(i)})| = \epsilon\} \cup \{x_{(n)}\}$, corresponding to points $\mathbf{x}_n$ where S_n touches the upper or lower boundary of the tube $T(E_n, \epsilon)$. The right derivative of S_n provides a histogram s_n on $[x_{(1)}, x_{(n)})$, which automatically determines the histogram bin number, bin locations and bin probabilities in (2). The taut-string s_n histogram is additionally known to have the fewest peaks or modes with an integral lying 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 (2004) prescribe a tube squeezing factor ϵ_n that determines the tube $T(E_n, \epsilon_n)$ and s_n as part of the TS histogram procedure. This is done using a data approximation concept involving weak metrics applied to a continuous distribution E and the empirical distribution E_n based on a sample from E. On these distributions, the κ-order Kuiper metric, $\kappa \in \mathbb{N}$, is given by

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

We define the difference between successive Kuiper metrics as $\rho_1(E, E_n) = d_{ku,1}(E, E_n)$ 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; let q_i denote the 0.999-quantile of $\rho_i(E, E_n)$. A definition can now be given for a taut-string to be consistent with the data $\mathbf{x}_n$. We say a taut-string distribution S_n from a tube $T(E_n, \epsilon)$ provides an adequate data approximation if $\rho_i(S_n, E_n) \le q_i$ for each $i = 1, \ldots, 19$.

In the taut-string (TS) procedure, we reduce the tube radius ϵ of $T(E_n, \epsilon)$ until the approximation standard is first met; this provides the squeezing factor ϵ_n to determine a final taut-string histogram s_n. Further squeezing beyond ϵ_n would create additional peaks or modes in a taut-string histogram.

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 percentages of silica in meteorites. The first two sets of data are found in Silverman (1985) and the last one in Good and Gaskins (1980). Extensive analyses by other authors have produced histograms with two modes for the Old Faithful data, right skewed histograms for the suicide study data [cf. Silverman (1985), Scott (1992)] and three modes of increasing size for the meteorite data [cf. Good and Gaskins (1980), Scott (1992)].
Figures 1-3 provide histograms with methods from Section 2. The point made visually is the degree to which histogram constructions disagree in their shapes, largely when it comes to the number and position of modes. We explore this aspect further in our numerical studies.

4 Simulation Study

Our simulation study focuses on the regular histogram procedures which were found to perform well by Birgé and Rozenholc (2006) as well as the taut-string method (TS). To limit the size of the study, we exclude several histogram procedures involving plug-in estimates, such as Sturges's rule of $1 + \log_2(n)$ bins [cf. Sturges (1926)], as well as methods from Daly (1988) and He and Meeden (1997). Numerical studies in Birgé and Rozenholc (2006) 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 a simulation study to compare performances of histogram construction procedures 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 $x \in I^\delta \backslash I$ for the enlargement $I^\delta = \cup_{y \in I}\{x \in \mathbb{R} : |x-y| \leq \delta\}$ 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 Scott (1992) for the normal density.

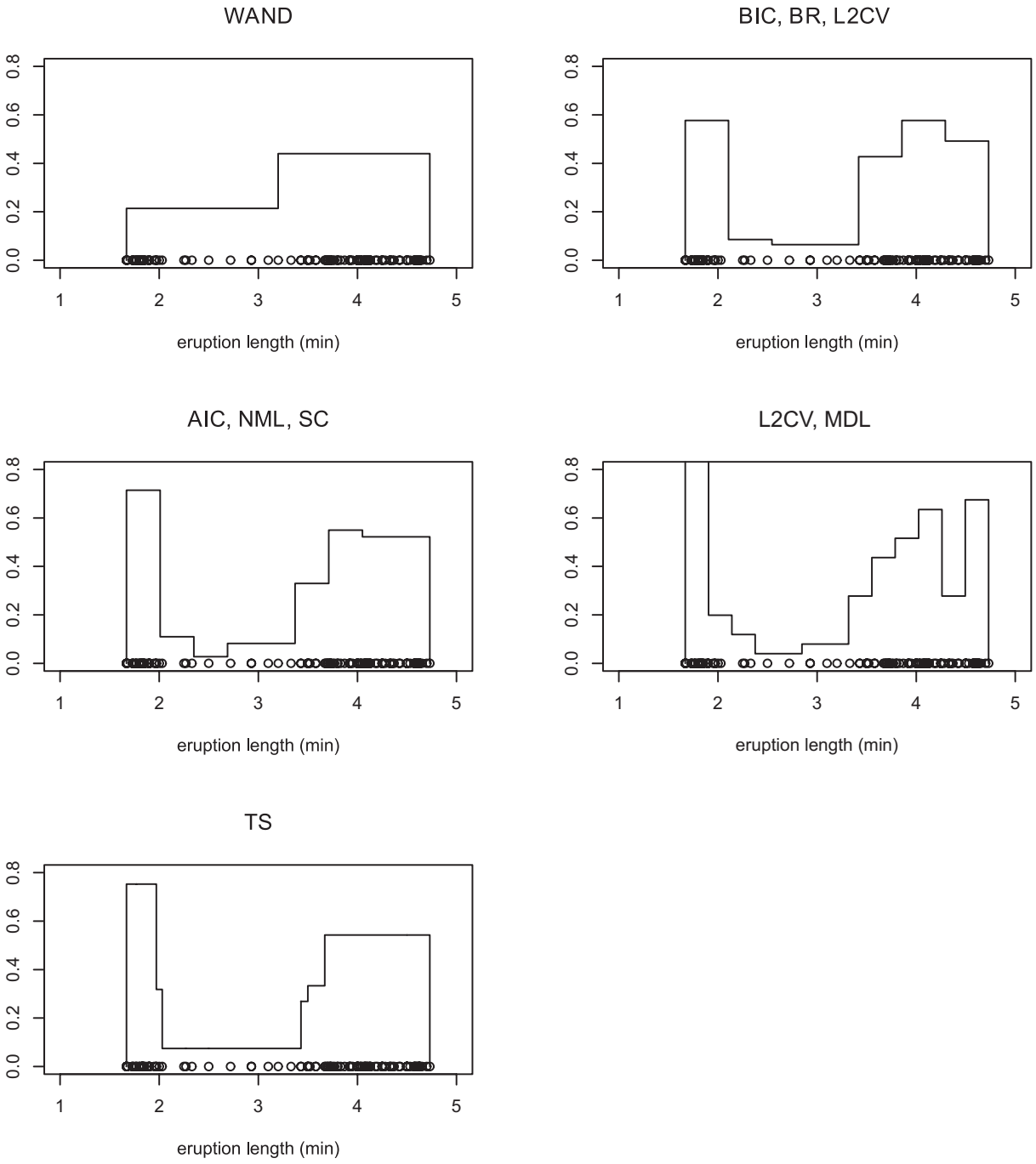


Figure 1: Histogram constructions from Old Faithful geyser data, $n = 107$ data points denoted by $\circ$.

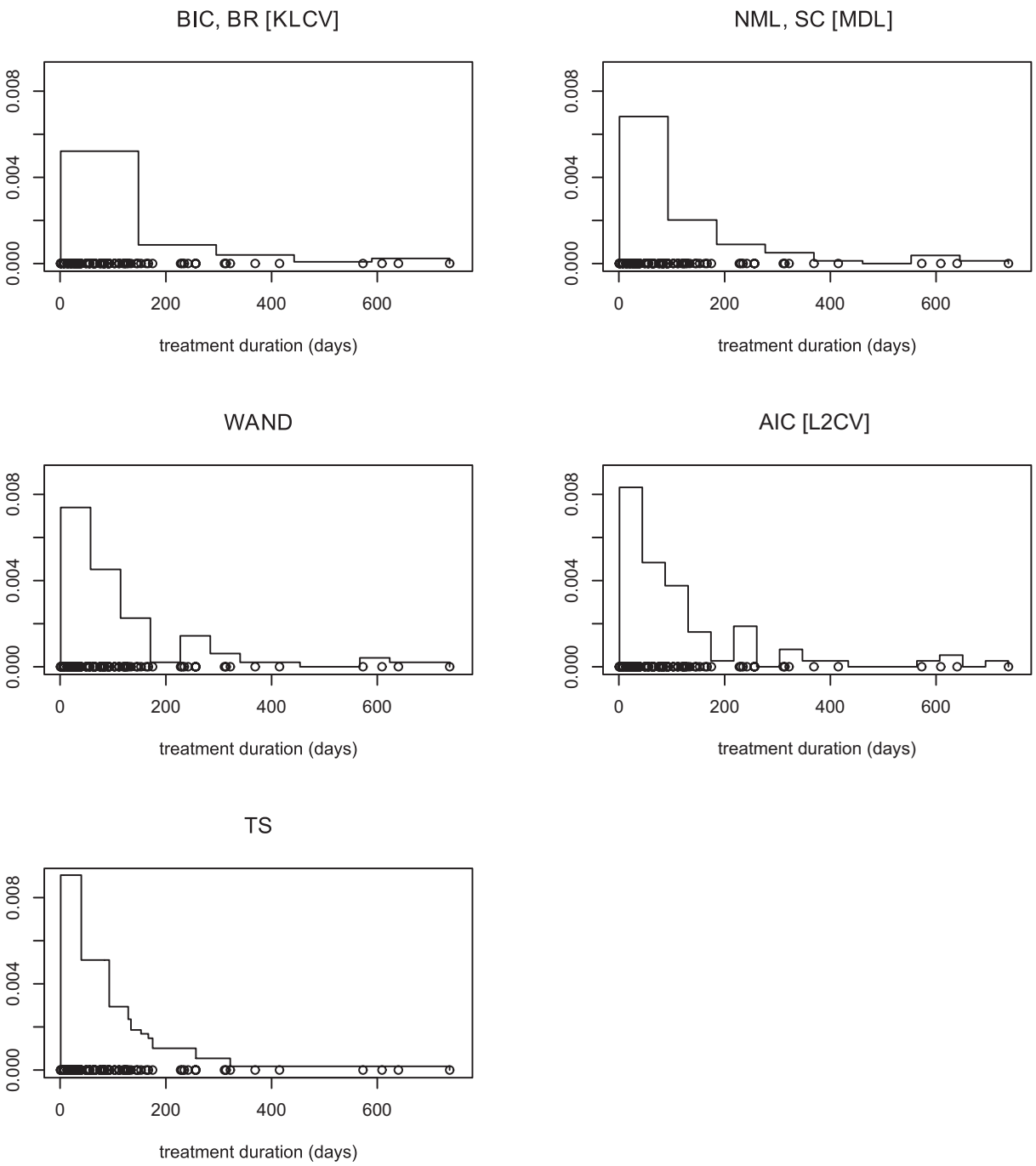


Figure 2: Histogram constructions from suicide data, $n = 86$. Methods in brackets $[\cdot]$ produced one additional bin compared to those histograms graphed and are similar in shape.

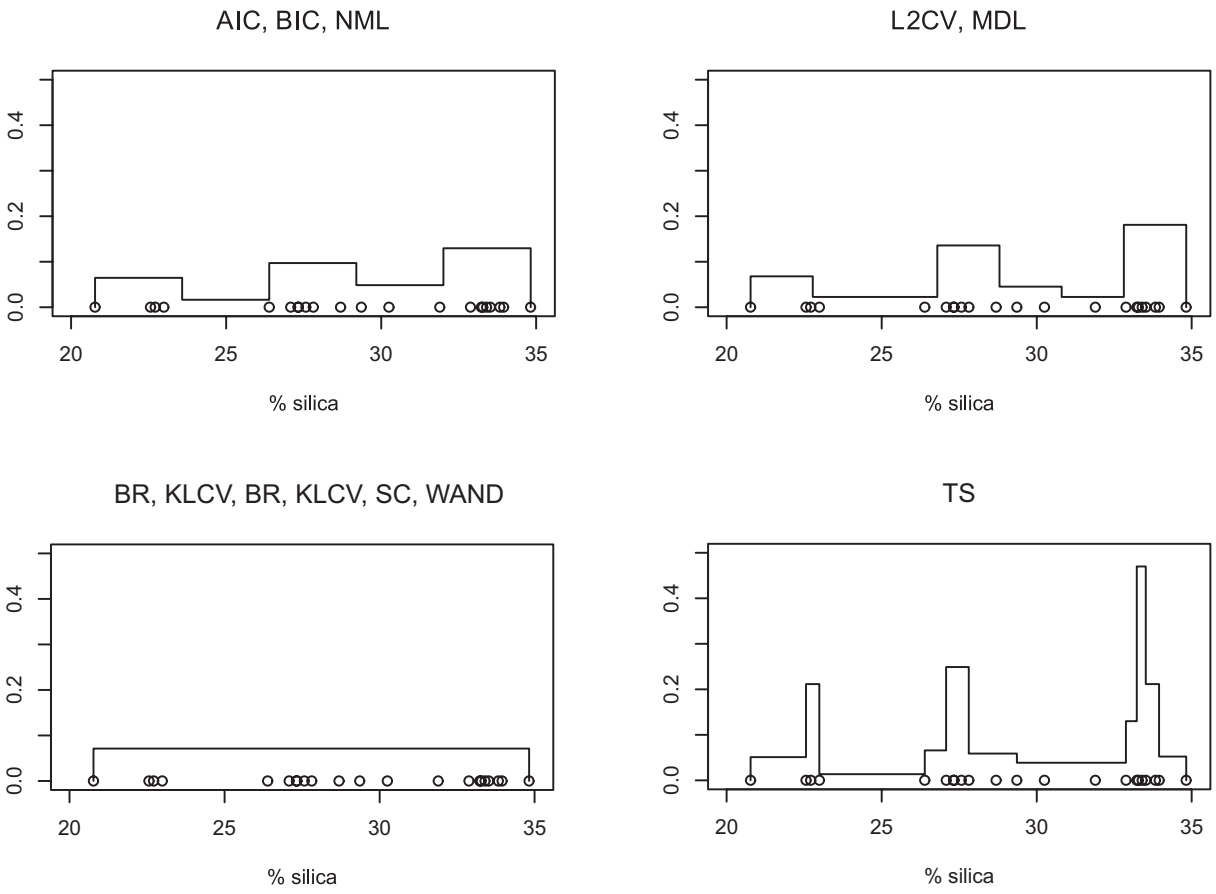


Figure 3: Histogram constructions from meteorite data, $n = 22$.

To illustrate this, Figure 4 provides histograms for a sample from the claw density, which is a normal mixture with five peaks taken from Marron and Wand (1992). 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 $z_1, \ldots, z_p$ satisfying $(z_i - \delta_i, z_i + \delta_i) \cap (z_j - \delta_j, z_j + \delta_j) = \varnothing$, $i \neq j$, for some positive vector $\delta = \delta(f) \equiv (\delta_1, \ldots, \delta_p) \in \mathbb{R}^p$. Assume next that a histogram $\hat{f}$ has $\hat{p} = \hat{p}(\hat{f})$ peaks at $y_1, \ldots, y_{\hat{p}}$. We say a peak of $\hat{f}$ at y_j matches a peak of f at z_i if $\min_{1 \leq j' \leq \hat{p}} |z_i - y_{j'}| = |z_i - y_j| < \delta_i$. An $\hat{f}$-peak 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:

$$\begin{aligned} \ell_{\text{i.d.}}(f, \hat{f}, \delta) &= \#\text{ of unidentified peaks of } f \;+\; \#\text{ of spurious peaks of } \hat{f} \\ &= (p - C_{\text{i.d.}}) + (\hat{p} - C_{\text{i.d.}}) \end{aligned} \tag{6}$$

using the number $C_{\text{i.d.}} = \sum_{i=1}^{p} \mathbb{I}\{\min_{1 \leq j \leq \hat{p}} |z_i - y_j| < \delta_i\}$ of correctly identified f-peaks. That is, the nonnegative loss $\ell_{\text{i.d.}}(f, \hat{f}, \delta) \geq 0$ measures the two possible errors incurred by identifying modes of f with the modes of $\hat{f}$. The vector δ represents the tolerances demanded in identifying each peak. Using $\ell_{\text{i.d.}}$ in (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 selected sixteen reference densities f of differing degrees of smoothness, tail behavior, support and modality. The collection of reference densities included: unimodal densities, such as the Uniform $U(0, 1)$, standard Normal $N(0, 1)$, and standard Cauchy distributions; eight mixture distributions from Marron and Wand (1992) as well as the claw density; a ten normal mixture in Figure 5 used by Loader (1999); and four remaining densities, which were chosen to have roughly the same shapes as the test-case densities appearing in Birgé and Rozenholc (2006) and have nearly all probability mass concentrated on $(0, 1)$.

The test densities are depicted in Figures 5 and 6.

We used these densities for evaluating the performance of eleven histogram procedures: AIC, BIC, BR, KLCV, L2CV, MDL, NML, SC, TS, WAND. To measure the quality of histograms, we considered risks based on two different losses: squared Hellinger and the peak identification loss from (6). 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 $n = (100, 250, 500)$, we used 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

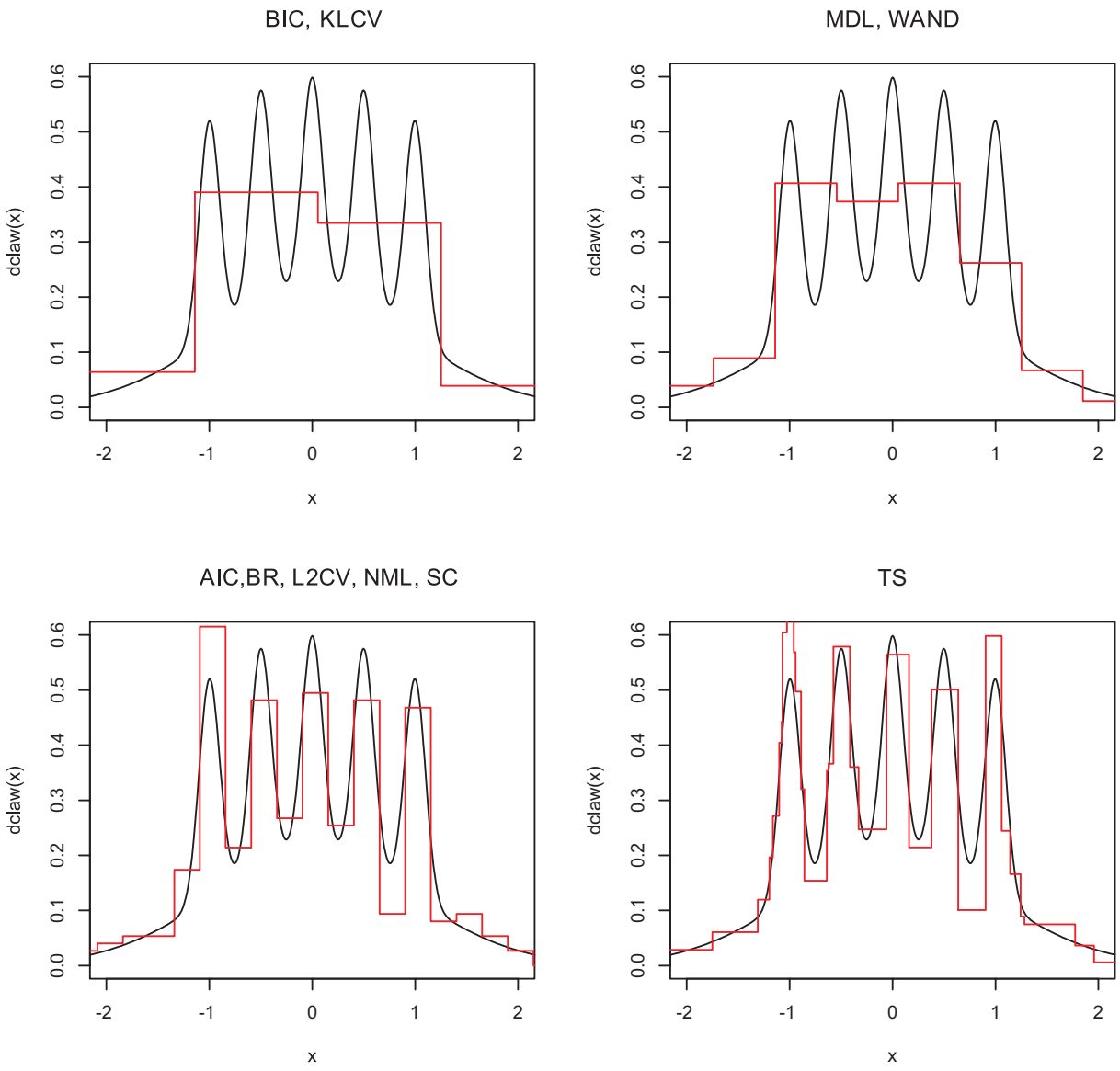


Figure 4: Histogram constructions for the claw density (black solid line) based on a sample $n = 300$.

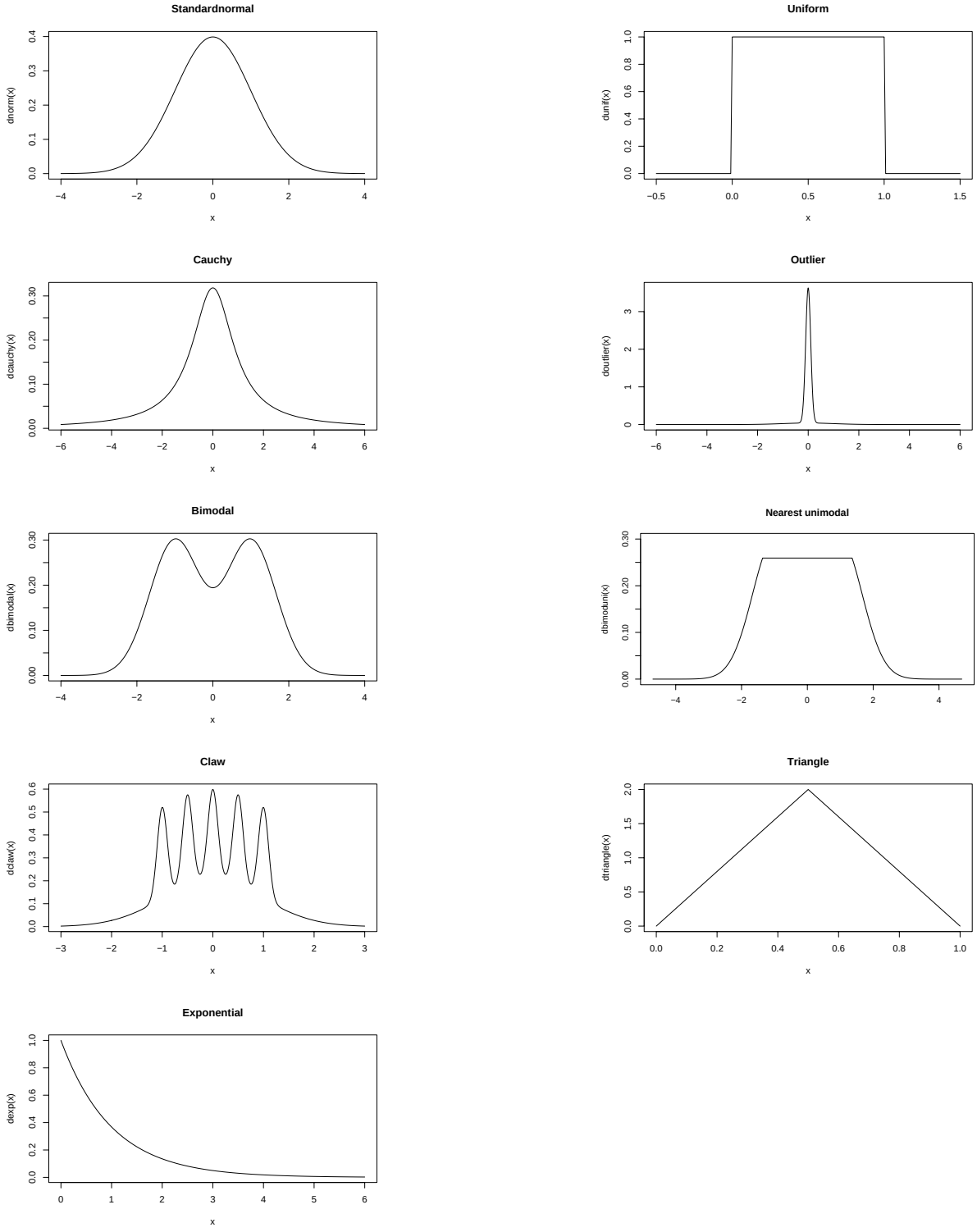


Figure 5: Data-generating densities used in the simulation study.

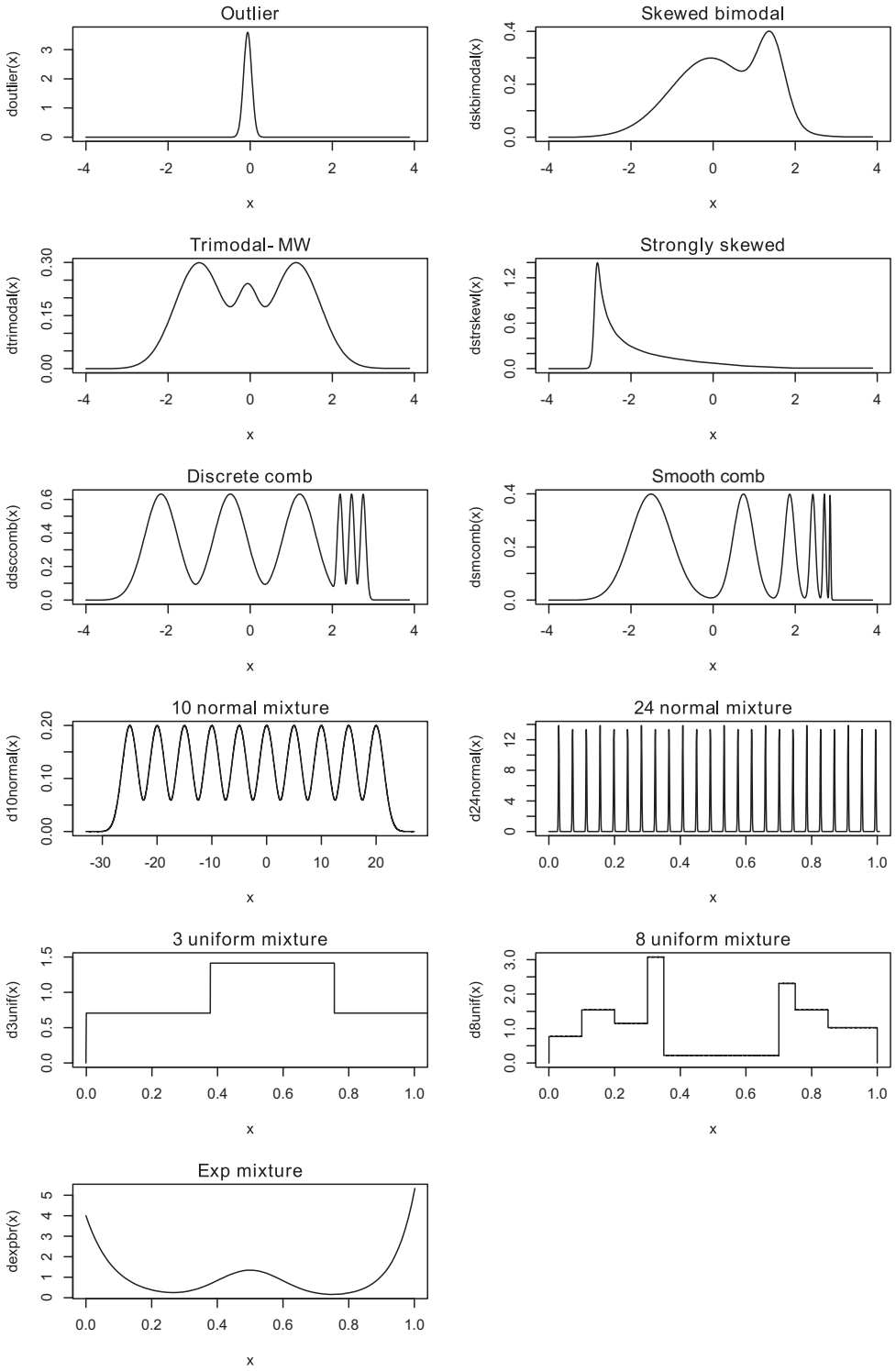


Figure 6: Data-generating densities used in the simulation study.

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 sample sizes $n = 100, 250, 500$.
In Figures 7 - 10 the average ranks of the ten methods are given. In Figures 7 and 9 the ranks are built over all test densities and all sample sizes of the simulation study, in Figures 8 and 10 the ranks are only taken over the unimodal or multimodal densities, respectively.
Sometimes care is needed in interpreting the results. An example is given by the results for the bimodel density. Here BIC performs best with a peak identification risk of only 0.09 for samples of size $n = 500$. This compares to 1.02 for the taut string. Moreover BIC performs better on the bimodal density than it does on the normal density where its risk is 0.19. The explanation is that most methods, including BIC for this density, find too many peaks and consequently tend to find some when there are none. 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 0.78 as against 0.03 for the taut string. The explanation is that the two peaks of the bimodal are not very pronounced and are difficult to find reliably. BIC seems to find them simply 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 of 1. In fact no method can detect these peaks reliably, BIC only gives the illusion of doing so.
We can now summarize the results of the simulation study as follows:

- The plug-in method (WAND) and L_2-cross-validation (L2CV) consistently were the worst performers in both peak identification and Hellinger distance.

- Out of the information theory-based hisograms, based on work of Rissanen (1987, 1989), the NML and SC performed similarly and were typically better than MDL. Agreement between NML and SC also appeared in the data examples (Figures 1-3).

- For Hellinger risk, the BR method did perform well in our simulations, but did not greatly outperform other histogram methods. In fact, on multimodal densities, there is not much difference indicated in Figure 10 among methods in terms of Hellinger risk.

- The taut-string procedure (TS) emerged as superior in terms of identifying the modes of a density. In this sense the TS histogram more accurately reflects the

shape of a density than the other methods. The TS was also comparable in terms of Hellinger distance.

As had to be expected, there is no overall optimal procedure that delivers the best histogram for every data-generating density. However, relative to any other single histogram method, the taut-string histogram provides good histogram approximations for a wide range of data-generating densities over a variety of sample sizes.

Acknowledgement

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).

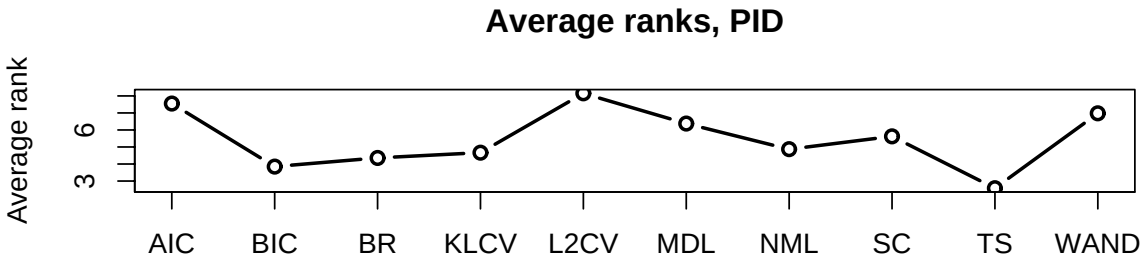
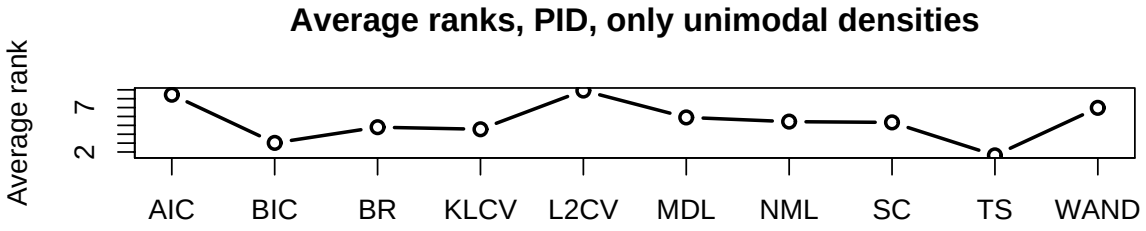


Figure 7: Average ranks for PID and all densities.



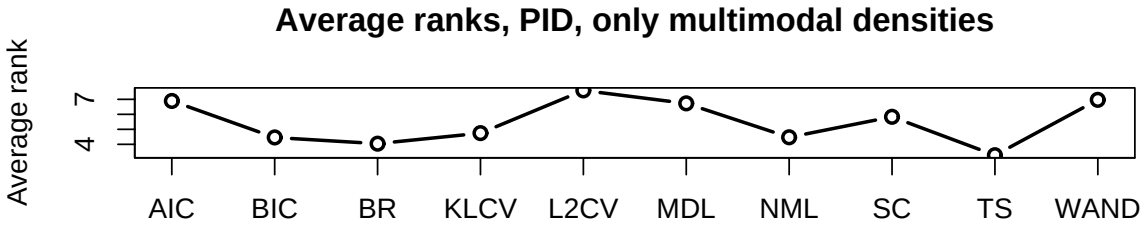


Figure 8: Average ranks for PID and unimodal or multimodal densities, respectively.

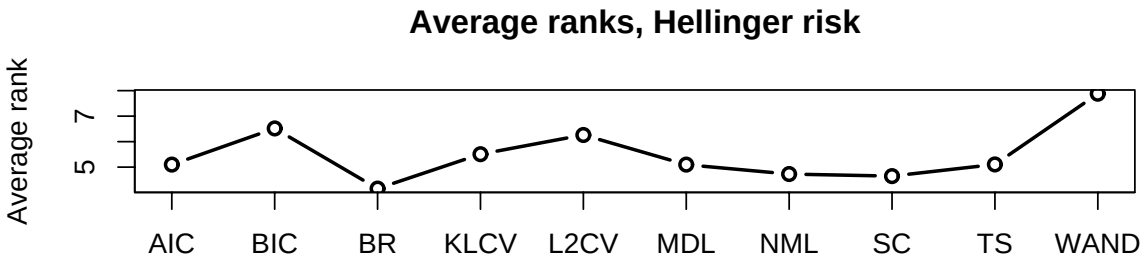
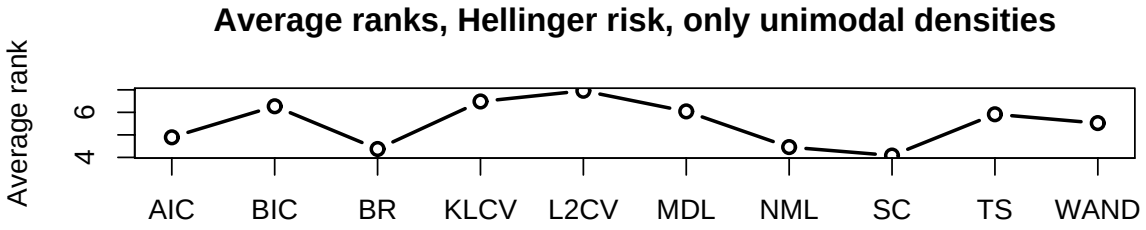


Figure 9: Average ranks for Hellinger loss and all densities.



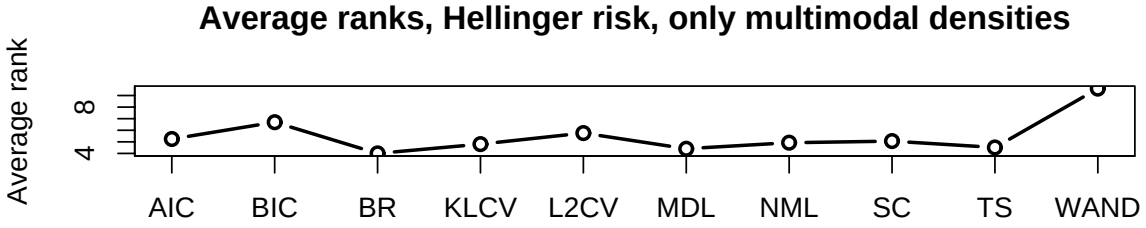


Figure 10: Average ranks for Hellinger loss and unimodal or multimodal densities, respectively.

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

density	n	AIC	BIC	BR	KLCV	L2CV	MDL	NML	SC	TS	WAND
$N(0,1)$	100	1.20	0.47	0.49	0.58	1.33	2.24	0.65	0.82	0.18	0.52
	250	1.17	0.34	0.37	0.43	1.38	1.61	0.40	0.44	0.03	0.37
	500	1.47	0.19	0.34	0.39	1.71	1.07	0.31	0.29	0	0.38
$U(0,1)$	100	0.72	0.02	0.07	0.41	0.57	6.11	0.03	0.03	0.58	1.25
	250	0.60	0.01	0.07	0.50	0.57	6.97	0.01	0	0.30	1.40
	500	0.55	0.01	0.06	0.52	0.53	2.45	0	0	0.32	1.54
cauchy	100	3.59	2.80	2.88	1.75	4.66	1.62	3.62	3.17	0.01	5.74
	250	5.85	4.37	4.65	1.89	8.25	1.89	5.76	5.10	0	11.56
	500	8.53	6.22	6.71	1.96	12.46	2.00	8.66	7.44	0	18.34
strongly skewed	100	3.75	2.68	2.84	2.48	5.30	3.11	3.28	3.25	0.81	2.00
	250	4.98	2.53	2.89	2.28	10.13	3.31	3.19	3.13	0.44	2.00
	500	8.55	2.46	4.07	2.05	13.65	3.64	3.78	3.79	0.23	1.97
outlier	100	3.64	2.08	2.41	1.39	5.03	1.14	3.45	2.46	0	4.98
	250	7.17	3.15	5.36	1.27	10.82	1.38	6.12	4.06	0	10.57
	500	10.72	4.87	9.04	1.25	17.44	1.71	9.03	6.72	0	18.00
three-uniform	100	1.36	0.07	0.10	0.40	1.36	3.78	0.27	0.76	0.12	0.90
	250	0.93	0	0.02	0.49	1.15	3.61	0.03	0.08	0.06	1.09
	500	0.67	0	0.01	0.47	0.94	0.90	0.01	0.01	0.08	1.54
bimodal	100	0.98	1.52	1.17	0.64	1.01	3.04	1.38	2.09	1.71	1.43
	250	0.80	0.25	0.22	0.32	0.85	2.16	0.21	0.32	1.53	0.59
	500	1.04	0.09	0.16	0.34	1.04	0.87	0.13	0.18	1.02	0.16
nearest unimodal	100	1.4	0.42	0.56	0.88	1.32	0.95	0.64	0.74	0.09	0.06
	250	1.64	0.79	0.94	1.15	1.62	1.06	0.96	1.07	0.04	0.2
	500	2.21	0.78	1.04	1.43	2.11	1.11	0.97	1.08	0.03	0.36
skewed bimodal	100	1.24	0.94	0.91	0.83	1.33	2.27	1.06	1.24	1.08	1.06
	250	1.41	0.67	0.64	0.69	1.49	1.77	0.70	0.71	1.01	0.95
	500	1.86	0.50	0.62	0.63	1.92	1.17	0.52	0.59	0.99	0.64
trimodal	100	1.62	1.94	1.77	1.42	1.59	3.34	1.80	2.23	2.21	2.13
	250	1.22	1.18	1.05	0.95	1.25	2.89	1.02	1.07	1.89	1.62
	500	1.22	0.91	0.73	0.75	1.20	1.26	0.77	0.72	1.48	1.25
exp mixture	100	1.82	1.55	1.33	3.14	3.36	1.51	1.78	1.42	0.60	3.94
	250	2.60	0.59	0.89	0.73	5.21	1.55	1.02	1.17	0.05	0.27
	500	3.32	0.37	1.17	0.82	6.69	3.03	0.81	1.13	0	0.06
eight-uniform	100	4.90	3.94	4.05	4.24	5.21	4.50	4.57	5.10	3.36	3.74
	250	4.83	3.86	3.99	4.08	5.26	4.70	4.06	4.34	2.35	4.52
	500	4.17	3.54	3.67	3.65	4.65	4.76	3.66	3.70	1.61	3.79
smooth comb	100	4.88	4.37	4.33	4.31	5.06	4.40	4.97	5.26	4.06	5.36
	250	5.88	3.98	3.84	4.01	5.81	3.73	4.42	4.81	3.10	5.52
	500	7.25	3.09	4.63	3.20	6.98	4.02	4.55	4.90	2.37	4.12
discrete comb	100	4.41	3.52	3.64	3.27	4.41	3.69	4.37	4.66	2.92	5.65
	250	4.45	3.87	3.74	3.86	4.76	3.78	3.96	4.14	2.09	5.66
	500	4.30	3.67	2.69	3.57	4.83	2.81	2.74	2.83	1.30	3.93
claw	100	5.18	5.46	5.49	5.40	5.01	4.99	5.39	5.33	4.25	5.28
	250	4.86	5.40	5.15	5.25	5.49	4.64	5.00	4.88	2.17	5.54
	500	5.51	5.35	4.12	5.02	7.03	4.24	4.23	4.16	0.36	5.57
ten-normal	100	4.89	10.90	10.16	10.94	5.01	6.07	7.37	6.60	6.20	10.48
	250	3.84	9.37	3.49	3.85	3.37	3.44	3.81	4.46	2.47	10.65
	500	3.09	4.11	2.19	2.44	2.68	2.64	2.26	2.70	1.06	11.44
24-normal	100	41.68	24.96	25.05	25.76	41.63	29.75	41.52	41.65	8.67	25.39
	250	23.48	33.78	39.88	30.43	24.14	30.70	23.66	23.49	2.41	25.80
	500	1.29	10.18	34.52	30.72	1.85	32.58	1.08	1.29	0.56	26.32
triangle	100	1.89	0.79	0.9	1.39	1.73	1.33	0.95	1.16	0.68	2.16
	250	1.85	0.89	1	1.55	1.78	1.15	1	1.18	0.46	2.29
	500	2.02	0.92	1.09	1.76	1.85	1.11	1.04	1.12	0.31	2.52
exponential	100	1.73	0.52	0.63	0.62	3.04	0.67	0.94	0.81	0	1.9
	250	2.3	0.7	1.07	0.63	4.15	0.75	1.17	0.96	0.02	3.82
	500	2.91	0.89	1.45	0.67	5.54	0.8	1.31	1.11	0.02	6.2

Table 2: Hellinger risk ($\times 100$) for histograms by density and sample size n.

density	n	AIC	BIC	BR	KLCV	L2CV	MDL	NML	SC	TS	WAND
$N(0,1)$	100	3.02	2.92	2.83	2.75	3.25	3.10	2.82	3.09	3.15	2.80
	250	1.47	1.63	1.48	1.48	1.55	1.47	1.49	1.45	1.93	1.43
	500	0.89	1.05	0.90	0.99	0.91	0.84	0.94	0.91	1.38	0.90
$U(0,1)$	100	1.72	1.05	1.11	1.38	1.55	4.10	1.00	1.04	1.40	1.32
	250	0.64	0.41	0.44	0.60	0.62	1.99	0.40	0.40	0.46	0.59
	500	0.31	0.21	0.22	0.30	0.30	0.47	0.21	0.20	0.25	0.34
cauchy	100	13.29	13.11	14.65	40.40	14.27	30.85	13.60	12.79	10.63	14.25
	250	11.17	11.82	13.73	50.33	12.10	40.77	11.02	10.79	8.62	12.09
	500	9.54	10.86	12.82	58.20	10.41	49.02	9.41	9.26	7.71	11.23
strongly skewed	100	4.44	3.98	3.93	3.95	5.31	3.60	4.26	4.01	3.80	10.37
	250	2.81	2.77	2.58	3.02	3.68	2.64	2.61	2.56	2.65	7.22
	500	2.06	2.38	2.00	2.93	2.31	2.23	2.03	2.01	1.66	5.98
outlier	100	6.33	6.69	6.54	24.53	6.35	14.14	6.33	6.48	3.92	6.10
	250	4.18	4.93	4.43	19.18	4.55	10.01	4.29	4.63	2.13	4.39
	500	2.91	3.79	2.94	16.10	3.46	8.00	2.95	3.30	1.46	3.45
three-uniform	100	2.77	1.82	1.84	2.02	2.65	3.32	2.05	2.37	2.17	3.54
	250	1.02	0.73	0.74	0.88	1.03	1.55	0.72	0.74	1.00	2.14
	500	0.46	0.36	0.36	0.43	0.48	0.47	0.37	0.36	0.53	1.52
bimodal	100	3.03	3.25	3.01	2.68	2.98	3.70	3.15	3.73	3.26	3.70
	250	1.50	1.66	1.52	1.44	1.57	1.71	1.53	1.51	1.82	2.05
	500	0.89	1.10	0.93	0.91	0.92	0.90	0.96	0.93	1.19	1.12
nearest unimodal	100	2.93	3.17	2.92	2.6	2.95	2.89	3.04	3.56	3.35	3.42
	250	1.43	1.55	1.43	1.38	1.49	1.38	1.44	1.41	1.88	2.27
	500	0.85	1.07	0.9	0.87	0.91	0.88	0.94	0.9	1.3	1.56
skewed bimodal	100	3.26	3.27	3.10	2.91	3.22	3.36	3.20	3.52	3.24	3.76
	250	1.65	1.77	1.63	1.68	1.69	1.61	1.64	1.61	1.98	2.48
	500	1.01	1.22	1.05	1.17	1.03	0.96	1.08	1.06	1.35	1.53
trimodal	100	5.76	5.96	5.76	5.42	5.70	6.50	5.90	6.34	5.91	6.40
	250	4.21	4.41	4.25	4.15	4.23	4.50	4.26	4.26	4.52	5.13
	500	3.62	3.84	3.66	3.61	3.64	3.62	3.69	3.65	3.79	4.36
exp mixture	100	5.35	4.61	4.60	4.18	5.86	3.70	5.16	5.24	4.05	5.96
	250	2.45	2.38	2.20	2.00	2.71	2.05	2.24	2.24	1.71	2.58
	500	1.36	1.49	1.30	1.20	1.47	1.38	1.30	1.29	0.96	1.41
eight-uniform	100	5.55	4.88	4.89	4.80	5.76	4.55	5.27	5.75	3.87	6.70
	250	2.32	3.25	2.36	2.71	2.39	2.17	2.30	2.21	1.71	4.53
	500	1.07	1.22	1.01	1.07	1.12	1.11	1.01	1.00	0.97	3.16
smooth comb	100	6.73	6.90	6.54	6.77	6.71	5.87	6.94	7.01	6.82	10.44
	250	3.92	4.39	3.89	4.36	3.93	3.65	3.87	3.80	3.69	8.16
	500	2.60	3.11	2.52	3.12	2.56	2.53	2.53	2.50	2.31	5.07
discrete comb	100	7.13	7.38	7.03	7.10	7.19	6.37	7.29	7.63	6.57	11.16
	250	4.22	4.50	4.17	4.63	4.25	3.94	4.19	4.17	3.68	8.88
	500	2.46	3.40	2.47	3.31	2.49	2.75	2.48	2.40	2.44	5.46
claw	100	5.14	4.22	4.17	4.32	5.51	4.00	4.68	4.98	4.51	5.47
	250	2.99	2.96	2.85	3.09	3.16	2.52	2.84	2.81	2.72	3.75
	500	1.91	2.33	1.79	2.49	2.03	1.85	1.85	1.82	1.60	2.90
ten-normal	100	8.65	7.05	7.13	7.40	8.26	6.50	8.80	9.39	8.74	7.19
	250	4.60	6.04	4.39	4.08	4.36	3.66	4.51	4.84	4.62	6.54
	500	2.76	3.55	2.68	2.67	2.64	2.33	2.71	2.69	2.77	6.41
24-normal	100	49.66	62.16	62.18	62.20	49.58	61.54	49.65	49.61	31.55	62.23
	250	21.66	30.15	51.86	60.57	21.99	60.58	21.66	21.67	15.25	61.82
	500	8.59	14.58	35.94	60.23	8.06	60.08	8.29	8.56	8.43	61.59
triangle	100	2.61	2.6	2.43	2.27	2.6	2.55	2.58	3.42	2.87	2.05
	250	1.29	1.4	1.29	1.2	1.31	1.28	1.31	1.28	1.51	1.04
	500	0.76	0.9	0.8	0.73	0.78	0.8	0.83	0.79	1.05	0.63
exponential	100	3.29	3.02	2.99	4.74	3.8	2.93	3.05	2.98	4.13	2.83
	250	1.68	1.68	1.61	3.76	1.96	1.85	1.61	1.61	2.65	1.69
	500	1.02	1.11	1	3.31	1.17	1.38	1.01	1.04	2.06	1.12

References

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

[2] Barron, A., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-413.

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

[4] Daly, J. E. (1988). Construction of optimal histograms. *Commun. Stat., Theory Methods* **17**, 2921-2931.

[5] Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution (with discussion). *Ann. Stat.* **29**, 1-65.

[6] Davies, P. L. and Kovac, A. (2004). Densities, spectral densities and modality. *Ann. Stat.* **32**, 1093-1136.

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

[8] Engel, J. (1997). The multiresolution histogram. *Metrika* **46**, 41-57.

[9] Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahr. verw. Geb.* **57**, 453-476.

[10] Good, I. J. and Gaskins, R. A. (1980). Density estimation and bump-hunting by the penalizes likelihood method exemplified by scattering and meteorite data. *Journal of the American Statistical Association*, **75**, 42-73.

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

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

[13] Hall, P. and Wand, M. P. (1988). Minimizing L_1 distance in nonparametric density estimation. *J. Multivariate Anal.* **26**, 59-88.

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

[15] Kanazawa, Y. (1988). An optimal variable cell histogram. *Commun. Stat., Theory Methods* **17**, 1401-1422.

[16] Kanazawa, Y. (1992). An optimal variable cell histogram based on the sample spacings. *Ann. Stat.* **20**, 291-304.

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

[18] Loader, C. R. (1999). Bandwidth selection: classical or plug-in? *Ann. Stat.* **27**, 415-438.

[19] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Stat.* **20**, 712-736.

[20] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **11**, 416-431.

[21] Rissanen, J. (1987). Stochastic Complexity (with discussion). *J. R. Statist. Soc. B* **49**, 223-239.

[22] Rissanen, J. (1989). *Stochastic complexity in statistical inquiry.* World Scientific, New Jersey.

[23] Rissanen, J., Speed, T. P. and Yu, B. (1992). Density estimation by stochastic complexity. *IEEE Trans. Inf. Theory* **38**, 315-323.

[24] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inf. Theory* **42**, 40-47.

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

[26] Schwartz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.

[27] Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66**, 605-610.

[28] Scott, D. W. (1992). *Multivariate density estimation: theory, practice, and visualization.* Wiley, New York.

[29] Silverman, B. W. (1978). Choosing the window width when estimating a density. *Biometrika* **65**, 1-11.

[30] Silverman, B. W. (1985). *Density estimation for statistics and data analysis.* Chapman and Hall, London.

[31] Szpankowski, W. (1998). On asymptotics of certain recurrences arising in universal coding. *Prob. Inf. Trans.* **34**, 142-146.

[32] Wand, M. P. (1997). Data-based choice of histogram bin width. *Amer. Stat.* **51**, 59-64.