Biometrika (2020), **107**, 2, *pp.* 347–364
Printed in Great Britain
doi: 10.1093/biomet/asz081
Advance Access publication 11 February 2020

The essential histogram

BY HOUSEN LI, AXEL MUNK, HANNES SIELING

Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidstr. 7, 37077 Göttingen, Germany

hli1@uni-goettingen.de munk@math.uni-goettingen.de hsielin@uni-goettingen.de

AND GUENTHER WALTHER

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, California 94305, U.S.A.

walther@stat.stanford.edu

SUMMARY

The histogram is widely used as a simple, exploratory way of displaying data, but it is usually not clear how to choose the number and size of the bins. We construct a confidence set of distribution functions that optimally deal with the two main tasks of the histogram: estimating probabilities and detecting features such as increases and modes in the distribution. We define the essential histogram as the histogram in the confidence set with the fewest bins. Thus the essential histogram is the simplest visualization of the data that optimally achieves the main tasks of the histogram. The only assumption we make is that the data are independent and identically distributed. We provide a fast algorithm for computing the essential histogram and illustrate our method with examples.

Some key words: Histogram; Mode detection; Multi-scale testing; Optimal estimation; Significant feature.

1. INTRODUCTION

The histogram, introduced by Karl Pearson in 1895, is one of the most basic, but still most widely used tools, for visualizing data. However, the construction of the histogram is not unique, leaving the user considerable freedom to choose the number and locations of breakpoints; see Freedman et al. (2007). This arbitrariness allows for radically different visual representations of the data, and it appears that no satisfactory rule for the construction is known, as evidenced by the large number of rules that have been proposed in the literature. In the case of equal bin width, popular examples of rules for the number of bins are those given by Sturges (1926), which is still the default rule in R, Scott (1979), Freedman & Diaconis (1981), Taylor (1987) and Birgé & Rozenholc (2006). Most of these rules are derived by viewing the histogram as an estimator of a density and choosing the number of bins to minimize an asymptotic risk estimate. This leads to questions about the performance for small samples as well as about smoothness assumptions that are not verifiable. Instead of having all bins of equal width, it is also common to specify equal area of all blocks. Denby & Mallows (2009) point out that the first approach typically leads to oversmoothing in regions of high density and is poor at identifying sharp peaks, whereas the second approach oversmooths in regions of low density and does not identify small

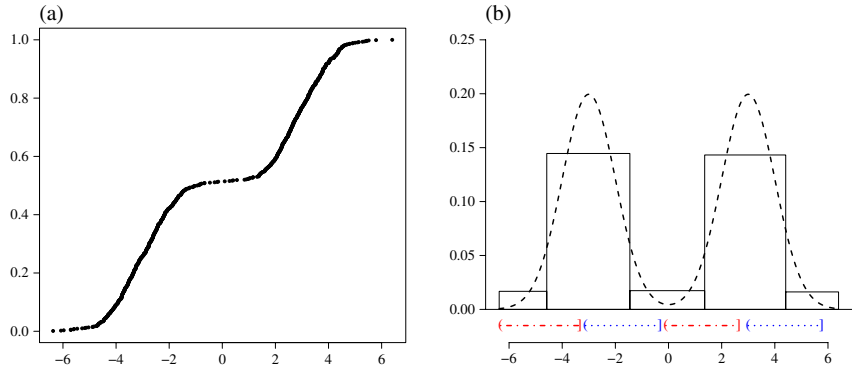

Fig. 1. Illustration of the essential histogram: (a) the empirical distribution of 900 observations from the Gaussian mixture $0.5N(-3,1)+0.5N(3,1)$; (b) the essential histogram with $\alpha = 0.1$ (solid) and the true density (dashed), with intervals of regions that contain a point of increase (dot-dash) or decrease (dotted) shown along the horizontal axis.

outlying groups of data. They advocate a compromise of these two approaches that is motivated by regarding the histogram as an exploratory tool for identifying structure in the data such as gaps and spikes, rather than as a density estimator, and they argue that relying on asymptotic risk minimization may lead to inappropriate recommendations for the number of bins. This is in line with recent findings for the regressogram (Tukey, 1961), the regression counterpart of the histogram (Frick et al., 2014; Li et al., 2016). Here bin choice corresponds to finding locations of constant segments, which is a different target from that of conventional risk minimization, e.g., of the L^p norm where $p \geqslant 1$.

This paper proposes a rule for constructing a histogram that is motivated by the two main goals of the histogram (see Freedman et al., 2007): to provide estimates of probabilities via relative areas, and to give a display of the density that is simple but informative, i.e., having few bins, but still showing important features of the data, such as modes.

The idea is to construct a confidence set of distribution functions such that each distribution function in the confidence set achieves the first goal in an asymptotically optimal way. Then, to attain the second goal, we select the simplest distribution function in the confidence set, i.e., the one with the fewest bins, as our histogram distribution function. The resulting histogram is the simplest one that shows important features of the data, such as increases and modes; we call this the essential histogram. Our approach is motivated by the fact that simplicity is a key feature of the histogram, implicit not only in the goal of having the histogram serve as an exploratory tool, but also in the definition of the histogram as a piecewise-constant function that should capture the major features of the data and the underlying distribution well. We show that in a large-sample setting, each distribution function in the confidence set estimates probabilities of intervals with a standardized simultaneous estimation error that is at most twice what is achievable and is typically much smaller than the errors obtained from histograms constructed with traditional rules. Likewise, we show that the distribution functions are asymptotically optimal for detecting important features, such as increases and modes of the distribution. Thus, the above two goals of the histogram are attained asymptotically. But one of the main benefits of our construction is that it provides finite-sample guaranteed confidence statements about features of the data: large increases of any histogram in the confidence set, and hence of the essential histogram, indicate significant increases in the true density, Theorem 3. We illustrate this with an example in Fig. 1. Our finite-sample guarantee ensures that the true density has an increase on the two dot-dash intervals and a decrease on the two dotted intervals in Fig. 1(b), with simultaneous confidence of at least 90%. This implies that the true density has two modes and one trough, as the plotted

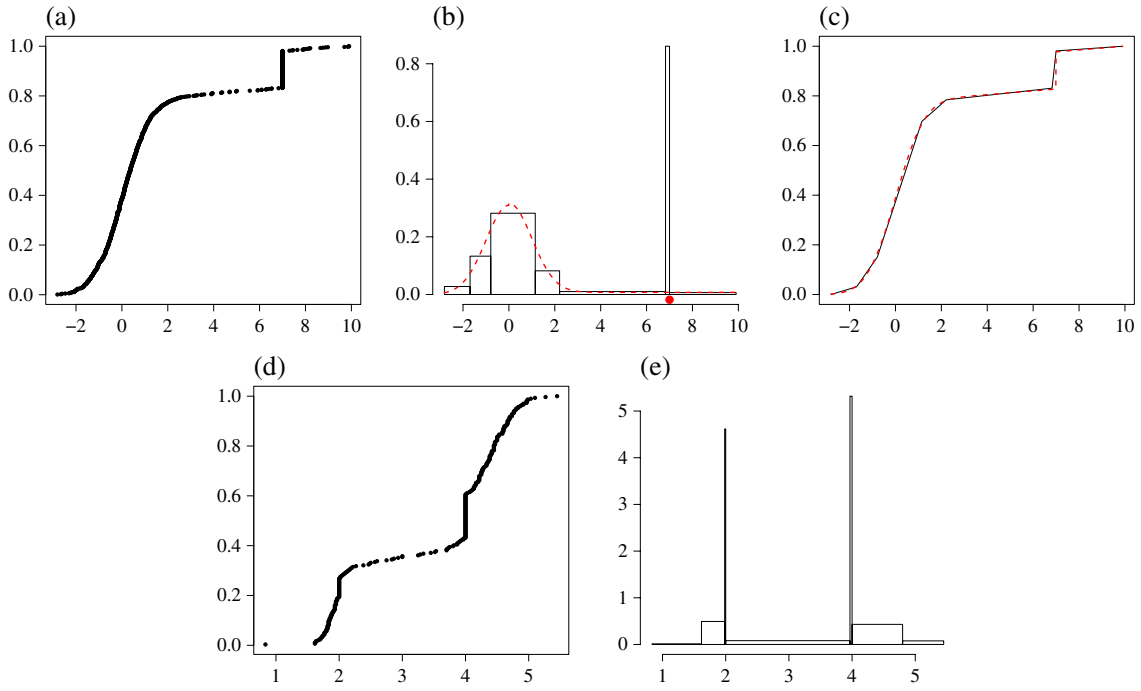

Fig. 2. Examples involving discontinuous distribution functions. The upper panels illustrate the Denby & Mallows (2009) example with sample size $n = 1000$: (a) the empirical distribution function; (b) the essential histogram with $\alpha = 0.5$ (solid) and the true density (dashed), with a dot indicating the point mass δ_7; (c) the distribution functions of the essential histogram (solid) and the truth (dashed). The lower panels illustrate the Geyser data example (Azzalini & Bowman, 1990): (d) the empirical distribution function of the durations in minutes; (e) the essential histogram with $\alpha = 0.5$.

intervals are disjoint (Dümbgen & Walther, 2008). These intervals are a selection from a much larger set of intervals of increase and decrease at all scales that the method provides; see § 3 and § 5. Thus, we can state with 90% guaranteed finite-sample confidence that these modes or troughs are really there in the underlying population. Such confidence statements are valuable enhancements of the essential histogram as an exploratory tool. Our method can be used in conjunction with any other histogram to obtain confidence statements for justifying or questioning the modes that the histogram suggests; see § 6.2. Indeed, the only parameter of the essential histogram is the significance level α. One should set α to 0.1 or even smaller if confidence statements are to be made, whereas one can take $\alpha = 0.9$ if the goal is to explore the data for potential features with tolerance to false positives. As a trade-off, we recommend $\alpha = 0.5$ as the default significance level.

The essential histogram is fairly general, since we make no assumption on F. In particular, it also applies to distributions with discrete components, which are common in real datasets (see, e.g., Unwin, 2015). Figure 2 gives two illustrative examples: one is the example in Denby & Mallows (2009, Fig. 4), which is a mixture of three distributions, $0.775\,N(0, 1) + 0.15\,\delta_7 + 0.075\,\mathrm{Un}(0, 10)$; the other consists of the time durations in the geyser dataset of Azzalini & Bowman (1990). As can be seen, the essential histogram estimates the probability of both continuous and discrete components over all scales rather well, and it reveals the true shape of the underlying distribution functions.

The construction of the confidence set is based on the multi-scale likelihood ratio test introduced by Rivera & Walther (2013), and we show here that this test results in the optimal detection of certain features in the data. Frick et al. (2014) employed such a multi-scale likelihood ratio test

for inference on changepoints in a regression setting, and they used the idea of selecting the function in the confidence set that has the fewest jumps. In the context of the histogram, this approach produces breakpoints only at locations where the evidence in the data requires them in order to show significant features and to provide good probability estimates. Hence the method will not put any breakpoints in regions where the density is close to being flat. This built-in parsimony is what one would expect from an automatic method of constructing a histogram; see also the comments about open research problems in Denby & Mallows (2009). The taut-string method of Davies & Kovac (2004) can be interpreted as producing a histogram that has the smallest number of modes within a confidence ball given by the periodic Kolmogorov metric, although the result does not meet the first goal of a histogram. It is known that the Kolmogorov metric will not result in good probability estimates for intervals unless they have large probability content (Dümbgen & Wellner, 2014). This procedure does not aim at parsimony of bins and will typically produce many more bins than the essential histogram, though often providing visually appealing solutions and estimating the number of modes well (see § 6), while the essential histogram automatically yields parsimony of bins and hence also of modes as explained above.

2. A CONFIDENCE SET FOR THE DISTRIBUTION FUNCTION

The empirical distribution function F_n of n independent and identically distributed univariate observations $X_1, \ldots, X_n$ is in a certain sense an optimal estimator of the underlying distribution function F; see Dvoretzky et al. (1956). While it is straightforward to convert F_n into a histogram distribution function (see Shorack & Wellner, 1986, p. 86), the resulting histogram with n breakpoints at the observations will generally not be useful for visualization of the data, as it is much too rough. The premise of this paper is that it is usually possible to remove a large proportion of these breakpoints, and still have an estimator that is just as good as F_n for estimating probabilities $F(I) = \int_I \mathrm{d}F$ of arbitrary intervals I. This is clearly plausible for local stretches where F has a density that is flat, but we will show that for more general F it is also typically possible to reduce the number of breakpoints considerably without incurring a significant error in estimating $F(I)$ or loss of power for detecting important features of F. This motivates our proposal to construct a histogram by choosing the histogram distribution function with the fewest breakpoints that is still optimal for the latter tasks. As the resulting histogram will typically be parsimonious, this construction achieves the goal of providing a simple visualization of the data that is optimally suited to the inferential and exploratory purposes of histograms.

The first step in this construction consists of deriving a confidence set of distribution functions that have the same performance as F_n in estimating probabilities $F(I)$. The idea is to apply certain likelihood ratio tests on a judiciously chosen set of intervals and then invert this family of tests, i.e., define a $(1-\alpha)$-confidence region for F as those distribution functions that pass the totality of these tests:

$$C_n(\alpha) = \Big\{\text{distribution function } H : \big[2 \log \mathrm{LR}_n\{H(I), F_n(I)\}\big]^{1/2} \leqslant \ell\{F_n(I)\} + \kappa_n(\alpha) \text{ for all } I \in \mathcal{J}\Big\}. \quad (1)$$

Here

$$\log \mathrm{LR}_n\{H(I), F_n(I)\} = nF_n(I) \log\left\{\frac{F_n(I)}{H(I)}\right\} + n\{1 - F_n(I)\} \log\left\{\frac{1 - F_n(I)}{1 - H(I)}\right\}$$

is the loglikelihood ratio statistic for testing $F(I) = H(I)$,

$$\ell\{F_n(I)\} = \left(2\log\left[\frac{e}{F_n(I)\{1-F_n(I)\}}\right]\right)^{1/2} \tag{2}$$

is the scale penalty, and $\kappa_n(\alpha)$ is the $(1-\alpha)$-quantile of the distribution of

$$T_n = \max_{I\in\mathcal{J}}\Big(\big[2\log \mathrm{LR}_n\{F(I), F_n(I)\}\big]^{1/2} - \ell\{F_n(I)\}\Big), \tag{3}$$

where $\mathcal{J}$ is a collection of intervals,

$$\begin{aligned} \mathcal{J} &= \bigcup_{l=2}^{l_{\max}} \mathcal{J}(l), \quad l_{\max} = \left\lfloor \log_2 \frac{n}{\log n} \right\rfloor \\ \mathcal{J}(l) &= \big\{(X_{(j)}, X_{(k)}] : j,k \in \{1+id_l,\, i\in\mathbb{N}_0\} \cap \mathcal{D},\, m_l < k-j \leqslant 2m_l\big\}, \end{aligned} \tag{4}$$

with

$$m_l = n\,2^{-l}, \quad d_l = \left\lceil \frac{m_l}{6l^{1/2}} \right\rceil, \quad \mathcal{D} = \{i : X_{(i)} \neq X_{(i+1)}\}.$$

This collection of intervals was introduced in Walther (2010) to approximate the collection of all intervals on the line in a computationally efficient manner; Rivera & Walther (2013) showed that the above multi-scale likelihood ratio statistic can be computed in $O(n\log n)$ steps, while the collection is still rich enough to guarantee optimal detection in certain scanning problems. In § 4 we show that every $H \in C_n(\alpha)$ has the same asymptotic estimation error as F_n for probabilities $F(I)$. Moreover, we show in § 5 that every $H \in C_n(\alpha)$ is optimal for the detection of certain features which are relevant for the exploratory purpose of the histogram. In particular, these optimality properties hold for the parsimonious histogram distribution function that we compute in § 3 in the second step of our construction of the essential histogram.

3. Computing the essential histogram

3.1. *Computationally feasible relaxation*

For a given partition of the real line into intervals $I_0, I_1, \ldots, I_K$, we define the histogram of F as the density $h(x) = \sum_{i=0}^{K} F(I_i)\mathbb{1}_{I_i}(x)/|I_i|$, where $|I_i|$ is the Lebesgue measure of I_i. The histogram h can be recovered from its distribution function H as the left-hand derivative of H. In the second step of our construction we will find a histogram in $C_n(\alpha)$ of (1) with the smallest number of bins. This computation requires the solution of a nonconvex combinatorial optimization problem and is practically infeasible for most real-world applications. However, it is possible to compute the exact solution of a slight relaxation, still nonconvex, of the original optimization problem in almost linear run-time; see § 3.2 and the Supplementary Material. This optimization problem is

$$\text{minimize } N_{\mathrm{bin}}(H) \quad \text{subject to} \quad H \in \tilde{C}_n(\alpha). \tag{5}$$

Here $\tilde{C}_n(\alpha)$ is the superset of the histogram distribution functions in $C_n(\alpha)$ that results if one evaluates the likelihood ratio tests only on those intervals where the candidate density is constant,

$$\tilde{C}_n(\alpha) = \Big\{\text{histogram distribution function } H : \big[2\log \mathrm{LR}_n\{H(I), F_n(I)\}\big]^{1/2} \leqslant \ell\{F_n(I)\} + \kappa_n(\alpha) \text{ for each } I \in \mathcal{J} \text{ where the left-hand derivative } H' \text{ is constant}\Big\}, \tag{6}$$

and $N_{\mathrm{bin}}(H)$ is the number of bins of the density of H. In general, solutions to (5) are not unique. In that case we will pick $\hat{H}$ with density $\hat{h} = \sum_{k=0}^{K} |I_k|^{-1} F_n(I_k) \mathbb{1}_{I_k}$, which maximizes the following negative entropy, up to a factor of n:

$$\sum_{k=0}^{K} nF_n(I_k) \log\left\{\frac{F_n(I_k)}{|I_k|}\right\}.$$

This is the loglikelihood if we assume that the data are distributed according to $\hat{H}$. Thus, among all solutions of (5) we select the one that explains the data best in terms of likelihood. We refer to this solution as the essential histogram.

Since $\tilde{C}_n(\alpha)$ is a superset of the histogram distribution functions in $C_n(\alpha)$, the minimization problem (5) over histogram distribution functions $H \in \tilde{C}_n(\alpha)$ will result in a solution that may have fewer bins than the minimizer over $C_n(\alpha)$, which is a beneficial side effect. In turn, $\tilde{C}_n(\alpha)$ involves fewer goodness-of-fit constraints, which may lead to some loss of efficiency in inference. In what follows, the theoretical results and the simulations show that this loss is not significant. Moreover, such computational relaxation still allows us to derive guaranteed finite-sample confidence statements about certain features of the distribution.

3.2. *Numerical computation*

For brevity, we focus on the main ideas here and defer the technical details to the Supplementary Material. The implementation of the numerics is provided in the `R` (R Development Core Team, 2020) package `essHist`, available on CRAN.

Computation of the threshold $\kappa_n(\alpha)$ in (1) is done in the following way. In the case of continuous F, the distribution of T_n is independent of F, so $\kappa_n(\alpha)$ can be determined by taking F to be, say, the uniform distribution, which leads to a confidence level that is exact at $1-\alpha$. In the general case, where F can be discontinuous, the distribution of T_n may depend on the unknown F. However, it is always stochastically bounded from above by a universal distribution, defined via a slight variant of T_n with F being uniform. In particular, this implies that there exists some $\kappa_n^*(\alpha)$ such that $\kappa_n(\alpha) \leqslant \kappa_n^*(\alpha)$ and $\sup_n \kappa_n^*(\alpha) < \infty$; see Lemma S2 in the Supplementary Material. This ensures that if one uses $\kappa_n^*(\alpha)$ instead of $\kappa_n(\alpha)$ as the threshold in (1), the confidence level is always at least $1-\alpha$, and all our theoretical results remain valid. The choice of $\kappa_n^*(\alpha)$, as compared to $\kappa_n(\alpha)$, makes the inference slightly conservative, but this is not consequential for the empirical performance of the essential histogram. In practice, we will choose $\kappa_n^*(\alpha)$ as the threshold in (1) when there are tied observations; otherwise, we treat F as continuous and estimate the threshold $\kappa_n(\alpha)$ by letting F be uniform. In all the experiments in this paper, $\kappa_n(\alpha)$ or $\kappa_n^*(\alpha)$ is estimated by 5000 Monte Carlo simulations, which needs to be done only once for a fixed sample size n and can be approximated for large n.

Computation of the essential histogram proceeds as follows. We denote by $X_{(1)}, \ldots, X_{(n)}$ the order statistics of the observations $X_1, \ldots, X_n$. We treat each $X_{(i)}$ as a node in a graph and define the edge length between nodes $X_{(i)}$ and $X_{(j)}$ to be the minimal number of blocks of a step function on $(X_{(i)}, X_{(j)}]$ that satisfies the multi-scale constraint (6). Then the computation of the essential

histogram involves finding the shortest path between $X_{(1)}$ and $X_{(n)}$, which can be computed exactly using dynamic programming algorithms (see, e.g., Dijkstra, 1959) with computation complexity $O(n^3)$. To improve computational speed, we exploit an accelerated dynamic program by incorporating pruning ideas (see, e.g., Killick et al., 2012; Frick et al., 2014; Hocking et al., 2017; Maidstone et al., 2017). The constraint that the estimator itself should be a histogram is incorporated into the dynamic programming algorithm. The resulting accelerated dynamic program is significantly faster than the standard dynamic program, and most of the time it has nearly linear computational complexity in terms of sample size, with the worst-case computational complexity being quadratic up to a log factor, which happens very rarely. This is confirmed by its empirical time complexity, which is almost linear. Moreover, the memory complexity is always $O(n)$.

4. Optimal estimation of probabilities

For ease of exposition, we assume here and in § 5 that the underlying distribution function F is continuous. We stress, however, that our method is designed to be able to deal with arbitrary and possibly discontinuous F. Recall that the distribution of T_n under any F is stochastically bounded from above by a universal distribution. Hence, the theoretical guarantees in Theorems 2, 3 and 5 carry over to any discontinuous F with natural modifications, and so does the upper bound in the first part of Theorem 1. In contrast, the lower bounds, in the second part of Theorem 1 and in Theorems 4 and S1, require the assumption of continuity on F in order to distinguish them from, for instance, the purely deterministic case. Some further optimality results and all the proofs are given in the Supplementary Material.

We now investigate how well $H \in C_n(\alpha)$ performs with regard to the first goal of the histogram, namely estimating probabilities $F(I)$ for intervals I. To this end, for probabilities of size $p \in (0, 1)$, we introduce the simultaneous standardized estimation error of H:

$$d_p(F, H) = \sup_{\text{intervals } I: F(I)=p} \frac{|H(I) - p|}{\{p(1-p)\}^{1/2}}. \tag{7}$$

Note that $d_p(F, H) = d_{1-p}(F, H)$. Therefore, it suffices to consider $p \in (0, 1/2]$ for $d_p(F, H)$.

The first result establishes a benchmark for this task by deriving the performance of the empirical distribution function F_n. It shows that $d_p(F, F_n)$ is very close to $\{2 \log(\mathrm{e}/p_n)/n\}^{1/2}$.

Theorem 1. *For $B_n \to \infty$ arbitrarily slowly as $n \to \infty$, we have that uniformly in F,*

$$\mathrm{pr}_F\left\{n^{1/2}\, d_p(F, F_n) \leqslant \left(2 \log \frac{\mathrm{e}}{p}\right)^{1/2} + B_n \text{ for all } p \in \left[\frac{\log^2 n}{n}, \frac{1}{2}\right]\right\} \to 1.$$

Furthermore, if $p_n \geqslant n^{-1} \log^2 n$ and $p_n \to 0$, then uniformly in F,

$$\mathrm{pr}_F\left\{n^{1/2}\, d_{p_n}(F, F_n) \geqslant \left(2 \log \frac{\mathrm{e}}{p_n}\right)^{1/2} - B_n\right\} \to 1.$$

In fact, no estimator can improve on the $\{2 \log(\mathrm{e}/p_n)/n\}^{1/2}$ bound, as explained in the proof of Theorem 1. Thus, F_n provides an optimal estimator for the collection $\{F(I)\}_I$. The next theorem shows that the distribution functions $H \in C_n(\alpha)$ nearly match this performance. The first part

says that if H_n is a fixed sequence of distribution functions such that $d_{p_n}(F, H_n)$ is slightly larger than this bound, then with high probability $H_n \notin C_n(\alpha)$. However, since we optimize over $C_n(\alpha)$ to find the simplest $H \in C_n(\alpha)$, we need to bound the worst-case estimation error over all $H \in C_n(\alpha)$. The second part of Theorem 2 shows that this worst-case error is at most twice the optimal bound. One can readily check that Theorem 2 holds also with $\tilde{C}_n(\alpha)$ in place of $C_n(\alpha)$ if in the definition of $d_p(F, H)$ we only consider intervals I where the density of H is constant.

THEOREM 2. *Let $B_n \to \infty$ and $\epsilon_n = B_n\{\log(\mathrm{e}/p_n\})^{-1/2}$. Then for $p_n \in (n^{-1}\log^2 n, 1/2)$ we have that uniformly in F,*

$$\sup_{H: d_{p_n}(F,H) > (1+\epsilon_n)\left(\frac{2}{n}\log\frac{\mathrm{e}}{p_n}\right)^{1/2}} \mathrm{pr}_F\{H \in C_n(\alpha)\} \to 0.$$

Moreover, uniformly in F,

$$\mathrm{pr}_F\left\{d_{p_n}(F,H) > (2+\epsilon_n)\left(\frac{2}{n}\log\frac{\mathrm{e}}{p_n}\right)^{1/2} \text{ for some } H \in C_n(\alpha),\ p_n \in \left(\frac{\log^2 n}{n},\ \frac{1}{2}\right)\right\} \to 0.$$

The loss of a factor of 2 is not consequential when compared to popular histogram rules: Proposition 1 gives the performance of a histogram that uses k_n equally sized bins. If one chooses $k_n \asymp n^{1/3}$ bins as recommended by the common rules in the literature, then $n^{1/2}d_{p_n}(F, H_n)$ blows up at the rate of $n^{1/3}$ for some quite typical continuous F and $p_n = (4k_n)^{-1}$, while the benchmark given by F_n and the worst-case error over $H \in C_n(\alpha)$ grow very slowly at a rate of $(\log n)^{1/2}$. A similar result is obtained if one uses bins with equal probability content.

PROPOSITION 1. *Let H_n denote the distribution function of a histogram that partitions $[0, 1]$ into k_n equally sized bins. Then there is a continuous F such that for $p_n = (4k_n)^{-1}$ and odd k_n,*

$$n^{1/2}\, d_{p_n}(F, H_n) \geqslant \frac{1}{2}(np_n)^{1/2}.$$

If one is willing to make higher-order smoothness assumptions on F, then it can be shown that the performance of these common histogram rules gets much closer to the benchmark. One key advantage of our proposed histogram is that it essentially attains the benchmark in every case by automatically adapting to the local smoothness. At the same time, some H in $C_n(\alpha)$ will typically have many fewer bins than the n produced by F_n: if the underlying density is locally close to flat, then the multi-scale likelihood ratio test will not exclude a candidate H that has no breakpoints in that local region. Thus the $H \in C_n(\alpha)$ with the fewest bins gives a simple visualization of the data while still guaranteeing essentially optimal estimation of $\{F(I)\}_I$.

The optimality results for estimating $F(I)$ in Theorem 2 and in Theorem S1 of the Supplementary Material carry over to estimating the average density $\bar{f}(I) = F(I)/|I|$ if one simply divides the inequalities by $|I|$; see § 5. The construction of $C_n(\alpha)$ via the loglikelihood ratio statistic $\mathrm{logLR}_n\{H(I), F_n(I)\}$ rather than, say, the standardized binomial statistic $n^{1/2}|H(I) - F_n(I)|\,[H(I)\{1 - H(I)\}]^{-1/2}$ is crucial for these optimality results; see the discussion in the Supplementary Material. That section also shows that $C_n(\alpha)$ is an optimal confidence region for F when $d_p(F, H)$ is interpreted as a distance between F and H.

5. Optimal detection of features

Besides estimating probabilities, another important purpose of a histogram is to show important features of the distribution, such as increases or modes of the density. An important aspect of the essential histogram is that the significance level of the confidence set $\tilde{C}_n(\alpha)$ automatically carries over to certain features of the essential histogram, thus making it possible to give finite-sample confidence statements about features of $\bar{f}(I) = F(I)/|I|$, which provides a measure of the average density over I without any smoothness assumptions on F. This is a noteworthy advantage of the essential histogram that is not shared by many other histogram rules. Such confidence statements about features of $\bar{f}$ can be derived from the following simultaneous confidence statement about $\bar{f}$.

Theorem 3. *Let* $c_n(I) = \ell\{F_n(I)\} + \kappa_n(\alpha)$ *with* $\ell\{F_n(I)\}$ *as in* (2)*, and let*

$$r_n(I) = \frac{2c_n(I)}{|I|}\left(\left[\frac{F_n(I)\{1-F_n(I)\}}{n}\right]^{1/2} + \frac{c_n(I)}{2n}\right).$$

Then with confidence of at least $1-\alpha$*,*

$$\left|\bar{f}(I) - \bar{h}(I)\right| \leqslant r_n(I) \tag{8}$$

simultaneously for all $I \in \mathcal{J}$ *and all* $H \in \tilde{C}_n(\alpha)$ *with density* h *constant on* I*.*

This simultaneous confidence statement can be used, for example, to establish finite-sample lower confidence bounds on the number of modes and troughs of $\bar{f}$. It follows from (8) that with confidence of at least $1-\alpha$, $\bar{f}(I) - \bar{f}(J)$ must have the same sign as $\bar{h}(I) - \bar{h}(J)$ whenever $|\bar{h}(I) - \bar{h}(J)| \geqslant r_n(I) + r_n(J)$. Therefore, if one can find intervals $I_1 < J_1 \leqslant I_2 < J_2 \leqslant \cdots \leqslant I_m < J_m$, where the inequalities are understood elementwise, such that $(-1)^{k+1}\{\bar{h}(I_k) - \bar{h}(J_k)\} > r_n(I_k) + r_n(J_k)$ for $k = 1, \ldots, m$, then one can conclude with confidence of at least $1-\alpha$ that $(-1)^{k+1}\{\bar{f}(I_k) - \bar{f}(J_k)\} > 0$, and hence that $\bar{f}$ has at least $\lfloor m/2 \rfloor + 1$ modes and $\lfloor m/2 \rfloor$ troughs. If F has density f, then $\bar{f}(I) - \bar{f}(J) > 0$ implies $f(x) > f(y)$ for some $x \in I$ and $y \in J$, so this confidence bound then applies to the density f as well. See Fig. 1 for an illustration.

We now show that the essential histogram is even optimal in reproducing such increases and decreases, in the sense that it will show an increase if the size of the increase in the underlying distribution is just above the threshold below which detection is asymptotically not possible. Since we are considering general distribution functions F and do not want to make any smoothness assumptions, we will quantify the size of an increase via $\bar{f}$. We consider a set $\mathcal{I}_n(c)$ of distribution functions that have an increase in $\bar{f}$ of size parameterized by $c > 0$:

$$\mathcal{I}_n(c) = \left\{F : \text{there exist disjoint intervals } I_1 < I_2 \text{ such that } \frac{\log^2 n}{n} < F(I_i) \leqslant p_n \text{ for } i = 1, 2 \right.$$
$$\left. \text{and } \bar{f}(I_2) - \bar{f}(I_1) > c \sum_{i=1}^{2} \frac{\left[2F(I_i)\{1-F(I_i)\}\log\frac{\mathrm{e}}{F(I_i)}\right]^{1/2}}{n^{1/2}|I_i|}\right\}, \tag{9}$$

where $p_n \in (2n^{-1}\log^2 n, 1/2)$ is any given sequence, which for simplicity we omit from the notation $\mathcal{I}_n(c) = \mathcal{I}_n(c, p_n)$. Theorem 4 shows that it is not possible to reliably detect an increase in $\bar{f}$ if $F \in \mathcal{I}_n(1-\epsilon_n)$ with $\epsilon_n \downarrow 0$ slowly enough, as no test to this effect can have nontrivial asymptotic power. In contrast, the first part of Theorem 5 says that with asymptotic probability 1,

the essential histogram will show the increase if $F \in \mathcal{I}_n(1+\epsilon_n)$. This result clearly also applies to the simultaneous reproduction of a finite number of increases/decreases and hence to the reproduction of modes. Thus the essential histogram has the desirable property that it will show increases and modes of $\bar{f}$ once the evidence in the data is strong enough to make the detection of these features possible in principle. Conversely, one needs to keep in mind that the presence of a feature such as an increase in the essential histogram does not automatically imply that the feature is present in $\bar{f}$; such an inferential confidence statement requires that the essential histogram show an increase that exceeds a certain size, as detailed in Theorem 3 and the subsequent exposition. The second part of Theorem 5 tells us that this condition is met if $F \in \mathcal{I}_n(3+\epsilon_n)$, losing only a factor of 3 on the optimal bound. This mirrors the result on the estimation of probabilities in Theorem 2, where a similar loss was found not to be consequential. Therefore, not only does the essential histogram have the advantage that it can provide confidence statements about certain features of $\bar{f}$, but when used as such an inferential tool it is even rate-optimal.

THEOREM 4. *Let $X_1, \ldots, X_n$ be independent samples from F, and write $X = (X_1, \ldots, X_n)$. Assume $\phi_n(X)$ is any test with level $\alpha \in (0,1)$ under the null hypothesis $H_0 : \bar{f}$ is nonincreasing in the sense that $\bar{f}(I_1) \geqslant \bar{f}(I_2)$ for all disjoint intervals $I_1 < I_2$. If $\epsilon_n \in (0,1)$ with $\epsilon_n(\log \mathrm{e}/p_n)^{1/2} \to \infty$, then*

$$\inf_{F \in \mathcal{I}_n(1-\epsilon_n)} E_F \phi_n(X) = \alpha + o(1).$$

THEOREM 5. *If $\epsilon_n > 0$ with $\epsilon_n(\log \mathrm{e}/p_n)^{1/2} \to \infty$, then*

$$\inf_{F \in \mathcal{I}_n(1+\epsilon_n)} \mathrm{pr}_F \Big\{ \text{every } H \in \tilde{C}_n(\alpha) \text{ whose density } H' \text{ is constant on } I_1 \text{ and } I_2 \text{ has a point of increase of } H' \text{ in the convex hull of } (I_1 \cup I_2) \Big\} \to 1.$$

If $\epsilon_n > 0$ with $\epsilon_n(\log \mathrm{e}/p_n)^{1/2} \to \infty$, then

$$\inf_{F \in \mathcal{I}_n(3+\epsilon_n)} \mathrm{pr}_F \Big\{ \text{for every } H \in \tilde{C}_n(\alpha) \text{ whose density } H' \text{ is constant on } I_1 \text{ and } I_2, \text{ the confidence statement (8) allows one to conclude } \bar{f}(I_2) > \bar{f}(I_1) \Big\} \to 1.$$

Furthermore, in the case where the underlying distribution is itself a histogram, i.e., has a piecewise-constant density, we have explicit control on the number of modes.

THEOREM 6. *Assume that the distribution function F has a piecewise-constant density $f = \sum_{k=0}^{K} c_k \mathbb{1}_{(\tau_k, \tau_{k+1}]}$, with $-\infty < \tau_0 < \cdots < \tau_{K+1} < \infty$. Then the essential histogram h with the distribution function H in (5) controls overestimation of the number of bins:*

$$\sup_F \mathrm{pr}_F \big\{ N_{\mathrm{bin}}(H) > N_{\mathrm{bin}}(F) \big\} \leqslant \alpha.$$

Furthermore, let

$$\gamma \equiv \gamma(f) = \lambda_f \min\{\underline{\theta}_f, \Delta_f\} \tag{10}$$

with $\underline{\theta}_f = \min_k c_k$, $\lambda_f = \min_k(\tau_k - \tau_{k-1})$ and $\Delta_f = \min_k |c_k - c_{k-1}|$, and assume that $\alpha_n \gtrsim n^{-\nu}$ for some $\nu > 0$ and that $\gamma_n \equiv \gamma(f_n) \geqslant c(\log n/n)^{1/2}$ for some small enough $c = c(\nu)$ and a

sequence of piecewise-constant densities $\{f_n\}_{n\geqslant 1}$ *with distribution functions* $\{F_n\}_{n\geqslant 1}$. *Then, for some generic constant* C, *the essential histogram* h *controls underestimation of the number of bins,*

$$\mathrm{pr}_{F_n}\{N_{\mathrm{bin}}(H_n) < N_{\mathrm{bin}}(F_n)\} \leqslant CK_n \exp(-Cn\gamma_n^2) \quad (n \geqslant n_0),$$

and controls the number of modes and troughs,

$$\mathrm{pr}_{F_n}(h_n \textit{ and } f_n \textit{ have the same number of modes and troughs}) \geqslant 1 - \alpha_n - CK_n \exp(-Cn\gamma_n^2) \quad (n \geqslant n_0).$$

In Theorem 6, the constants c, C and n_0 are known explicitly; see Proposition S1 in the Supplementary Material. Therefore, a sufficient condition for the consistent estimation of the number of bins, modes and troughs is

$$\gamma_n \gtrsim \frac{b_n + (\log K_n)^{1/2} + (\log n)^{1/2}}{n^{1/2}}$$

for some $b_n \to \infty$, which can be arbitrarily slow. Further, we stress that γ in (10) quantifies the underlying difficulty in estimating the numbers of modes and troughs and the number of bins.

6. Simulation study

6.1. *Comparison study*

In this section we consider various simulation scenarios that reflect a range of difficulties in density estimation and data exploration. For comparison, we include the classical histograms with bins of equal width (Pearson, 1895) or blocks of equal area (Scott, 1992), as well as a more recent multi-scale density estimator proposed by Davies & Kovac (2004). The number of bins for the classical histograms is selected by the rule of Sturges (1926), the default in `R`, and the asymptotically optimal rule of Scott (1992). Both are computed with the built-in `R` function `hist`. The Davies–Kovac estimator has a similar flavour to the essential histogram, defined as a solution to a variational problem under a certain multi-scale constraint, but it computes only an approximate solution using taut strings together with some heuristic adjustments, such as local squeezing, so it is difficult to make statistical error guarantees or confidence statements. It is computed using the function `pmden` with default parameters in the `R` package `ftnonpar`, available on CRAN. We report only visual results here; detailed comparisons in terms of mean integrated squared error, skewness, and number of modes can be found in the Supplementary Material.

Scenario 1: Uniform density. Observations are from the uniform distribution $\mathrm{Un}(0, 1)$. The results of the comparison are shown in Fig. 3 and the Supplementary Material. The essential histogram performs best with small significance levels, $\alpha \leqslant 0.5$, recovering the true density almost perfectly, whereas with large α, such as $\alpha = 0.9$, like the Davies–Kovac estimator it tends to include false bins. In sharp contrast, the classical histograms perform worst overall and have many false bins $\{N_{\mathrm{bin}}(\hat{F}) - N_{\mathrm{bin}}(F)\}$ and hence false modes, and the performance becomes even worse as the sample size increases.

Scenario 2: Monotone density. Figure 4 and the Supplementary Material show results for the exponential distribution with unit mean. The essential histogram is better than the other methods

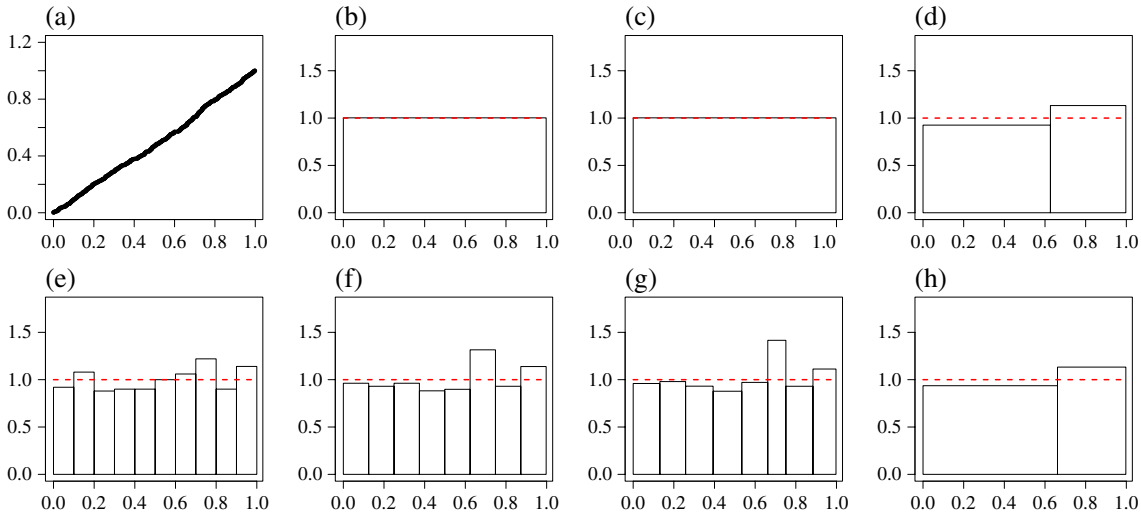

Fig. 3. Uniform density: (a) the empirical distribution function; (b), (c), (d) the essential histograms with $\alpha = 0.1, 0.5, 0.9$, respectively; (e), (f) the histograms with bins of equal width selected using the rules of Sturges (1926) and Scott (1992); (g) the histogram with blocks of equal area according to the rule of Scott (1992); (h) the Davies–Kovac estimator. In each panel, the true density is represented by a dashed line and the sample size is $n = 500$.

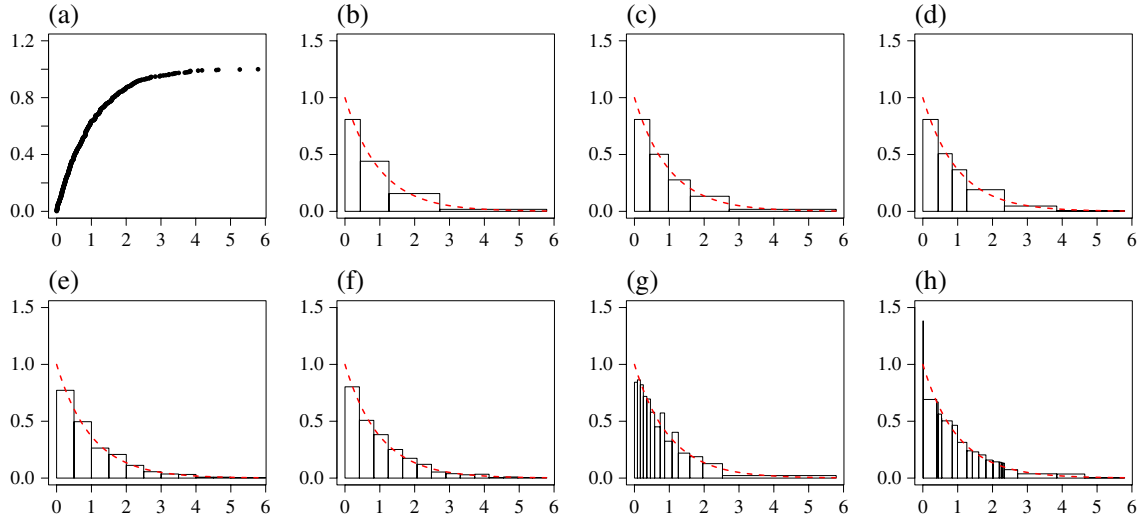

Fig. 4. Exponential density: see Fig. 3 for the descriptions of panels (a)–(h). The sample size is $n = 500$.

from both density estimation and feature detection perspectives, while requiring the fewest bins, which eases the interpretation of the data. The Davies–Kovac estimator performs similarly well, but sometimes distorts the true shape, as with the artificial spike in Fig. 4(h). As in the previous example, the classical histograms are less competitive and tend to include more false modes as the sample size increases. The comparison results on other monotone densities, not shown, are similar to this example.

Scenario 3: Histogram density. The distribution in this example is

$$\tfrac{1}{4}\mathrm{Un}(0,2) + \tfrac{1}{8}\mathrm{Un}(0.75,1.25) + \tfrac{1}{8}\mathrm{Un}(2.975,3.025) + \tfrac{1}{2}\mathrm{Un}(4,6),$$

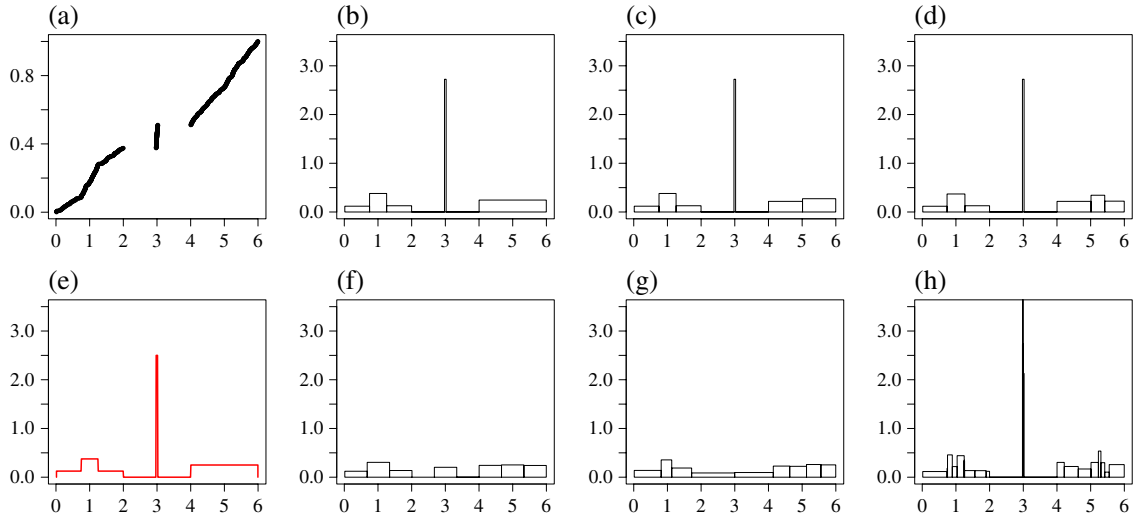

Fig. 5. Histogram density: see Fig. 3 for the descriptions of panels (a)–(d) and (f)–(h); panel (e) shows the true density. The sample size is $n = 800$.

which consists of three different regions: an ordinary one-mode region, a sharp-spike region, and a flat region. The results of the comparison are given in Fig. 5 and in the Supplementary Material. The essential histogram with a wide range of significance levels performs substantially better than all the other methods. It recovers all three regions of the true density fairly well and greatly outperforms the other methods with respect to detection of the correct number of bins, which confirms the theoretical finding in Theorem 6. For a fixed sample size, the essential histogram tends to introduce slightly more false bins and slightly more false modes at larger significance levels α. By contrast, the Davies–Kovac estimator often noticeably overestimates the height of the spike and introduces many distinct modes in the one-mode and flat regions. Its ability to identify the true number of modes first improves, but then deteriorates as the sample size increases. The classical histograms again perform worst and seriously flatten the central spike.

Scenario 4: Claw density. The results of the comparison on the claw density from Marron & Wand (1992) are shown in Figs. 6 and 7 and in the Supplementary Material. The essential histogram performs well in terms of both mode detection and density estimation for large sample sizes n or at high significance levels α. For a fixed n, it recovers more details of the density from the data as α increases, at the expense of statistical confidence. This illustrates the usefulness of the essential histogram as a potential exploratory tool for the analysis of data, and we suggest viewing the nominal level α as a screening parameter. A small α provides reliable confidence statements in Theorems 3 and 5; a large α typically leads to better recovery, such as in mode detection. For a fixed α, the performance of the essential histogram improves as n increases, which supports the theoretical finding in Theorem 5. Also, the essential histogram needs the fewest bins to detect the correct number of modes. Empirically, solutions in a range of α between 0.5 and 0.9 always look very similar, see Fig. 6, revealing a certain stability if estimation is the primary goal. Moreover, the essential histogram recovers the shape of the truth in such a reliable way that the skewness of the estimated histograms almost coincides with that of the truth. The Davies–Kovac estimator is among the best for mode detection, but it slowly starts to include more false modes as n increases. It does not perform as well in estimating the height of each mode, and the number of bins within each peak varies to a great extent; see Fig. 6(h). The latter behaviour could lead to misinterpretation of the data; for example, one might wrongly infer that the peaks are of

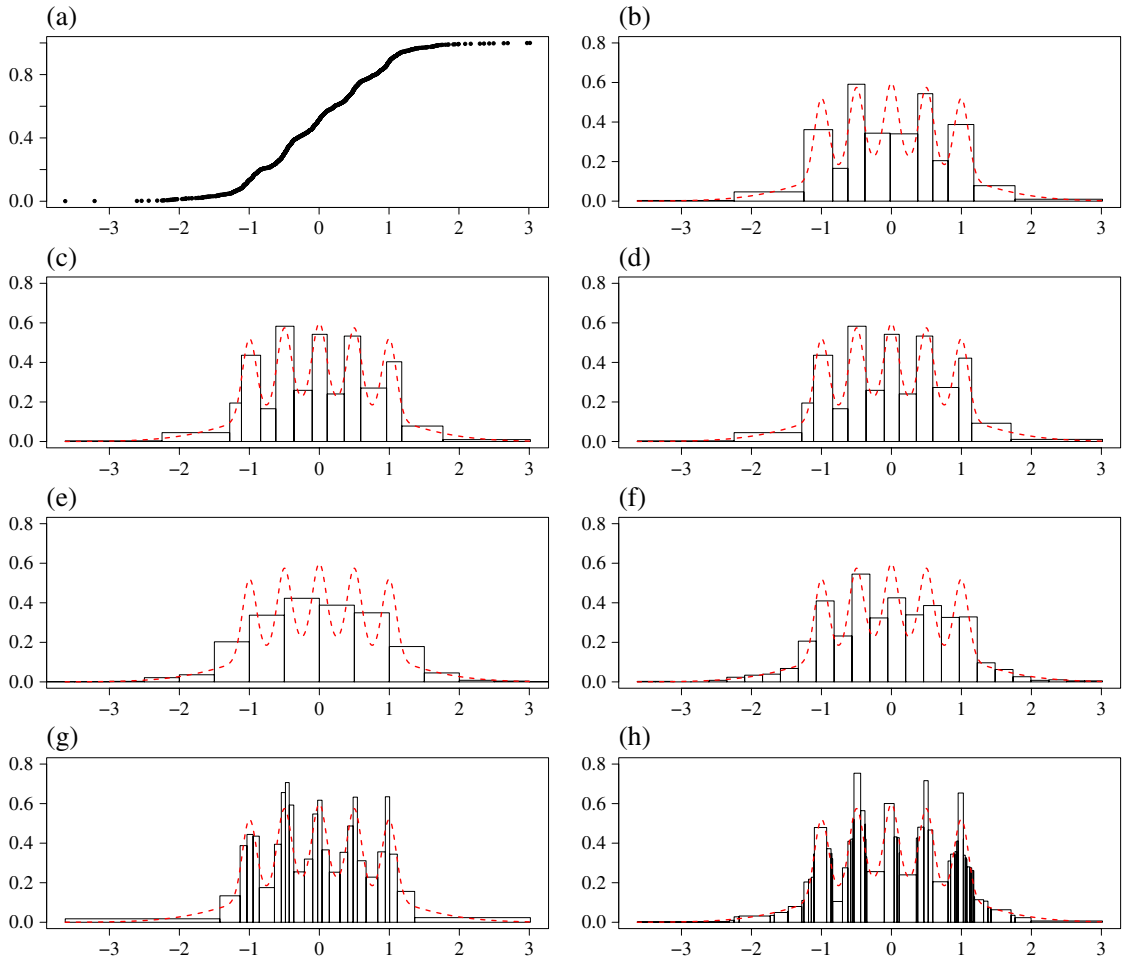

Fig. 6. Claw density: see Fig. 3 for the descriptions of panels (a)–(h). The sample size is $n = 1500$.

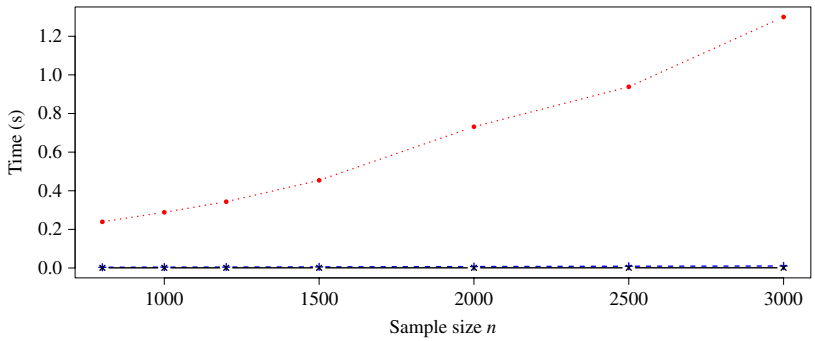

Fig. 7. Computation time, averaged over 500 runs, in the claw density example on a laptop computer with a single 3.3 GHz processor, two cores and 8 GB of memory, comparing the classical histograms (solid), the Davies–Kovac estimator (dashed) and the essential histogram (dotted).

completely different shape. In addition, the Davies–Kovac estimator gives the largest number of bins among all the methods, which further complicates interpretation of the data. For the classical histograms, Scott's rule is better than the rule of Sturges in terms of both mode detection and skewness preservation, but it tends to report more modes, both true and false, as n increases. The

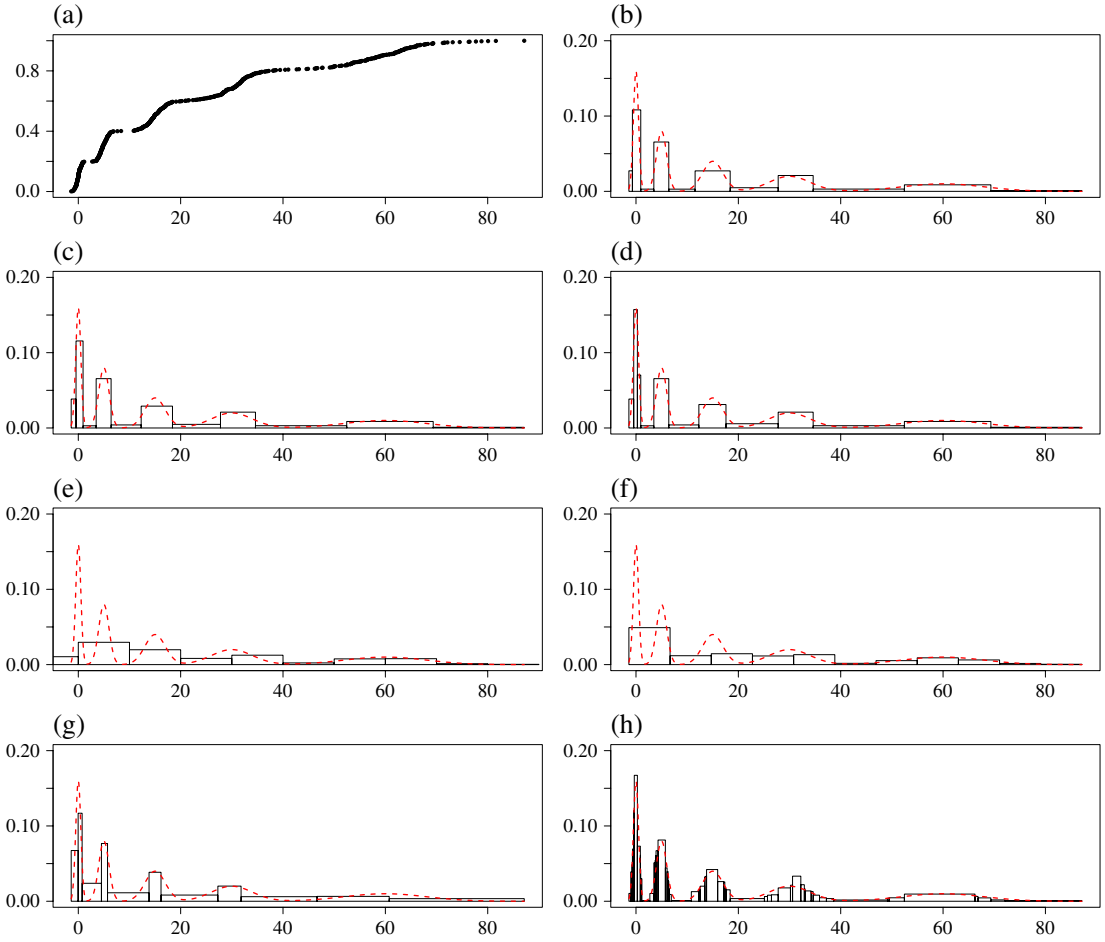

Fig. 8. Harp density: see Fig. 3 for the descriptions of panels (a)–(h). The sample size is $n = 800$.

equal-bin-width histogram gives better estimation in the tail region, with low density, while the equal-block-area histogram is superior in the central region, with high density. With respect to computation time, the essential histogram is the slowest while still being affordable; for example, it takes only around 1 second for 3000 observations; see Fig. 7. It appears that the computation time is of the same order for all the methods, that is, linearly increasing in n.

Scenario 5: Harp density. This example considers the Gaussian mixture density $0.2N(0, 0.5) + 0.2N(5, 1) + 0.2N(15, 2) + 0.2N(30, 4) + 0.2N(60, 8)$, termed the harp density due to the resemblance of its shape to a harp; see Fig. 8. It has modes at several scales, which are increasingly more difficult to detect from left to right. The results of the comparison are shown in Fig. 8 and in the Supplementary Material. The essential histogram with various significance levels α is the best overall at recovering the shape of the true density, as can be seen in Fig. 8(b)–(d), and it is also quantitatively the best at reproducing the skewness. Concerning mode detection, the essential histogram with larger α usually performs better, at the expense of lower confidence about the inference. For large sample sizes, $n \geqslant 1500$, the essential histogram with different α will eventually identify the correct number of modes. Further, the essential histogram outperforms all the other methods in terms of estimation error measured by the Kolmogorov metric, and is only slightly worse than the Davies–Kovac estimator in terms of the mean integrated squared error. The Davies–Kovac estimator is again the best at mode detection, but it has a tendency to bias the

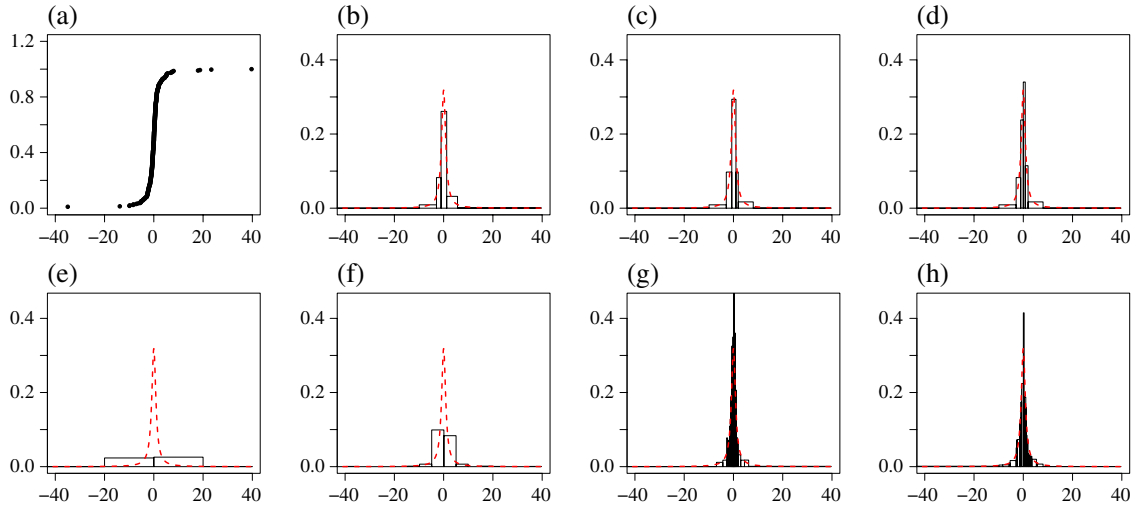

Fig. 9. Cauchy density: see Fig. 3 for the descriptions of panels (a)–(h). The sample size is $n = 300$.

exact shapes and locations of modes; see, for instance, the local maxima near 30 in Fig. 8(h); it also significantly underestimates the skewness of the truth. The classical histograms are generally less competitive; visually, the equal-bin-width histograms perform better in the region $[40, 60]$, while the equal-block-area histogram is better in $[0, 40]$; see Fig. 8(e)–(g). Moreover, the equal-block-area histogram is preferred for mode detection and lowering estimation error, but the equal-bin-width histogram is more favourable for skewness preservation. This dilemma in deciding between these two types of histogram reflects the underlying difficulty of the problem.

Scenario 6: Heavy tails. This comparison is performed on the standard Cauchy density $f(x) = 1/\{\pi(1 + x^2)\}$, a typical density with heavy tails. The results are shown in Fig. 9 and in the Supplementary Material. Overall, the essential histogram and the Davies–Kovac estimator outperform the classical histograms. Both perform almost perfectly in mode detection. For density estimation, the essential histogram recovers the truth quite well with only a few bins, while the Davies–Kovac estimator tends to include many unnecessary slim bins and sometimes overestimates the peak of the truth. Further, the essential histogram is most robust against outliers, as indicated by the little changes in the number of bins. The classical histogram with bins of equal width detects the major features, but can substantially overestimate the true peak. By contrast, the classical histogram with blocks of equal area completely distorts the shape of the truth, although it still identifies the correct number of modes with moderate frequency.

6.2. *Multi-scale constraint as an evaluation tool*

The multi-scale constraint $\tilde{C}_n(\alpha)$ in (5) can be helpful if used with any histogram estimator $\hat{\mu}$ as an evaluation tool. For example, for each I in $\mathcal{J}$ on which $\hat{\mu}$ is constant, we can check whether the corresponding local constraint $[2 \log \mathrm{LR}_n\{\int_I \hat{\mu}(x)\,\mathrm{d}x, F_n(I)\}]^{1/2} - \ell\{F_n(I)\} \leqslant \kappa_n(\alpha)$ is fulfilled. The collection of all intervals where the local constraints are violated shows whether and where $\hat{\mu}$ misses important features, i.e., false negatives. Further, $\tilde{C}_n(\alpha)$ can be used to find superfluous breakpoints, i.e., false positives. To this end, we consider whether merging two nearby estimated segments will still satisfy $\tilde{C}_n(\alpha)$. If so, the breakpoint there is said to be removable. Although the removable breakpoints cannot all be removed simultaneously, any subcollection of removable breakpoints such that any two are not endpoints of a common segment are simultaneously

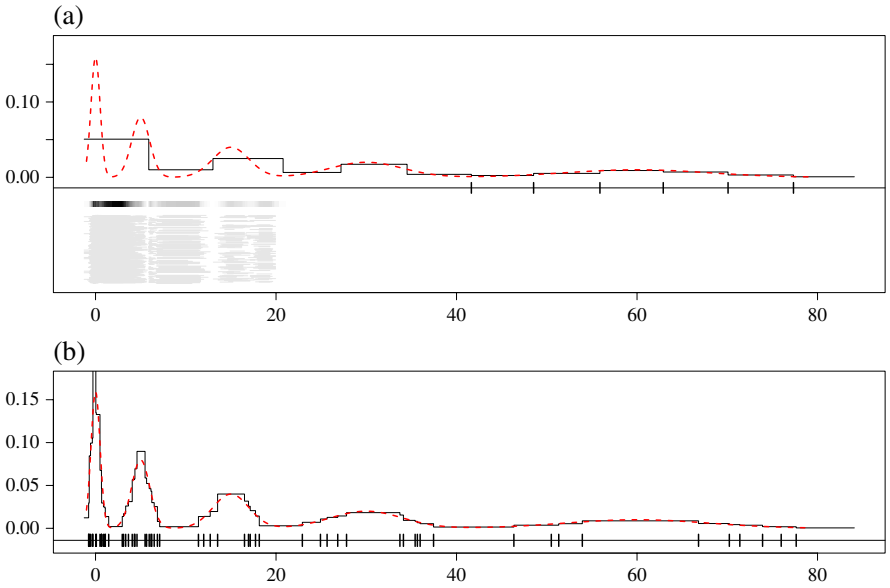

Fig. 10. Evaluation for the harp example: (a) the equal-bin-width histogram with Scott's rule; (b) the Davies–Kovac estimator. In each panel, the sample size is $n = 10^3$ and the truth is represented by a dashed line; the lower part shows intervals where violation of local constraints occurs. The grey-scale bars in panel (a) summarize the violations of local constraints, with the darkness indicating the number of violation intervals covering a location. The short vertical lines mark removable breakpoints.

removable. The evaluation in terms of violation intervals and removable breakpoints is simultaneously valid with confidence $1 - \alpha$; see Theorems 3 and 5. An example is shown in Fig. 10: the classical histogram recovers modes of medium size well, but it misses the spiky modes and reports redundant breakpoints for the widely spread modes; the Davies–Kovac estimator, on the other hand, has no violation intervals and hence misses no modes, but it gives many unnecessary breakpoints.

Acknowledgement

We thank the editors and reviewers for constructive feedback, as well as Anthony Unwin for helpful comments and for pointing us to several datasets which led to an improvement of our method. The first three authors were supported by the Deutsche Forschungsgemeinschaft, and the fourth author was supported by the U.S. National Science Foundation.

Supplementary material

Supplementary material available at *Biometrika* online includes all proofs, computational details and results of additional simulations. The `R` package `essHist` available on CRAN implements the proposed method.

References

Azzalini, A. & Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Appl. Statist.* **39**, 357–65.

Birgé, L. & Rozenholc, Y. (2006). How many bins should be put in a regular histogram. *ESAIM Prob. Statist.* **10**, 24–45.

Davies, P. L. & Kovac, A. (2004). Densities, spectral densities and modality. *Ann. Statist.* **32**, 1093–136.

DENBY, L. & MALLOWS, C. (2009). Variations on the histogram. *J. Comp. Graph. Statist.* **18**, 21–31.
DIJKSTRA, E. W. (1959). A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–71.
DÜMBGEN, L. & WALTHER, G. (2008). Multiscale inference about a density. *Ann. Statist.* **36**, 1758–85.
DÜMBGEN, L. & WELLNER, J. (2014). Confidence bands for distribution functions: A new look at the law of the iterated logarithm. *arXiv:* 1402.2918.v2.
DVORETZKY, A., KIEFER, J. & WOLFOWITZ, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.* **27**, 642–69.
FREEDMAN, D. & DIACONIS, P. (1981). On the histogram as a density estimator: L_2 theory. *Z. Wahr. verw. Geb.* **57**, 453–76.
FREEDMAN, D. A., PISANI, R. & PURVES, R. A. (2007). *Statistics*. New York: W. W. Norton & Co., 4th ed.
FRICK, K., MUNK, A. & SIELING, H. (2014). Multiscale change point inference. *J. R. Statist. Soc.* B **76**, 495–580. With 32 discussions by 47 authors and a rejoinder by the authors.
HOCKING, T. D., RIGAILL, G., FEARNHEAD, P. & BOURQUE, G. (2017). A log-linear time algorithm for constrained changepoint detection. *arXiv:* 1703.03352.
KILLICK, R., FEARNHEAD, P. & ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Am. Statist. Assoc.* **107**, 1590–8.
LI, H., MUNK, A. & SIELING, H. (2016). FDR-control in multiscale change-point segmentation. *Electron. J. Statist.* **10**, 918–59.
MAIDSTONE, R., HOCKING, T., RIGAILL, G. & FEARNHEAD, P. (2017). On optimal multiple changepoint algorithms for large data. *Statist. Comp.* **27**, 519–33.
MARRON, J. S. & WAND, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20**, 712–36.
PEARSON, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Phil. Trans. R. Soc.* A **186**, 343–414.
R DEVELOPMENT CORE TEAM (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
RIVERA, C. & WALTHER, G. (2013). Optimal detection of a jump in the intensity of a Poisson process or in a density with likelihood ratio statistics. *Scand. J. Statist.* **40**, 752–69.
SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66**, 605–10.
SCOTT, D. W. (1992). *Multivariate Density Estimation*. New York: John Wiley & Sons.
SHORACK, G. R. & WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. New York: John Wiley & Sons.
STURGES, H. A. (1926). The choice of a class interval. *J. Am. Statist. Assoc.* **21**, 65–6.
TAYLOR, C. C. (1987). Akaike's information criterion and the histogram. *Biometrika* **74**, 636–9.
TUKEY, J. W. (1961). Curves as parameters, and touch estimation. In *Proc. 4th Berkeley Sympos. Math. Statist. Prob.*, vol. I. Berkeley, California: University of California Press, pp. 681–94.
UNWIN, A. (2015). *Graphical Data Analysis with R*. Boca Raton, Florida: Chapman and Hall/CRC.
WALTHER, G. (2010). Optimal and fast detection of spatial clusters with scan statistics. *Ann. Statist.* **38**, 1010–33.

[*Received on* 27 *January* 2017. *Editorial decision on* 20 *August* 2019]