Digital Signal Processing 95 (2019) 102581

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Optimal data-based binning for histograms and histogram-based probability density models

Kevin H. Knuth

Dept. of Physics, Univ. at Albany (SUNY), Albany, NY 12222, USA

ARTICLE INFO

Article history:
Available online 11 September 2019

Keywords:
Binning data
Histogram
Probability density
Density models
Small sample size
Sufficient data

ABSTRACT

Histograms are convenient non-parametric density estimators, which continue to be used ubiquitously. Summary quantities estimated from histogram-based probability density models depend on the choice of the number of bins. We introduce a straightforward data-based method of determining the optimal number of bins in a uniform bin-width histogram. By assigning a multinomial likelihood and a non-informative prior, we derive the posterior probability for the number of bins in a piecewise-constant density model given the data. In addition, we estimate the mean and standard deviations of the resulting bin heights, examine the effects of small sample sizes and digitized data, and demonstrate the application to multi-dimensional histograms.

1. Introduction

Histograms are used extensively as non-parametric density estimators both to visualize data and to obtain summary quantities, such as the entropy, of the underlying density. However in practice, the values of such summary quantities depend on the number of bins chosen for the histogram, which given the range of the data dictates the bin width. The idea is to choose a number of bins sufficiently large to capture the major features in the data while ignoring fine details due to 'random sampling fluctuations'. Several rules of thumb exist for determining the number of bins, such as the belief that between 5 and 20 bins is usually adequate (for example, `Matlab` uses 10 bins as a default). Scott [1,2] and Freedman and Diaconis [3] derived formulas for the optimal bin width by minimizing the integrated mean squared error of the histogram model $h(x)$ of the true underlying density $f(x)$,

$$L(h(x), f(x)) = \int dx \big(h(x) - f(x)\big)^2. \tag{1}$$

For N data points, the optimal bin width v goes as $\alpha N^{-1/3}$, where α is a constant that depends on the form of the underlying distribution. Assuming that the data are normally distributed with a sample variance s gives $\alpha = 3.49s$ [1,2], and

$$v_{\text{scott}} = 3.49 s N^{-1/3}. \tag{2}$$

Given a fixed range R for the data, the number of bins M then goes as

$$M_{\text{scott}} = \lceil \frac{R}{3.49s} N^{1/3} \rceil. \tag{3}$$

Freedman and Diaconis report similar results, however they suggest choosing α to be twice the interquartile range of the data. While these appear to be useful estimates for unimodal densities similar to a Gaussian distribution, they are known to be suboptimal for multimodal densities. This is because they were derived by assuming particular characteristics of the underlying density. In particular, the result obtained by Freedman and Diaconis is not valid for some densities, such as the uniform density, since it derives from the assumption that the density f satisfies $\int f'^2 > 0$.

Another approach by Stone [4] relies on minimizing

$$L(h, f) - \int f^2$$

to obtain a rule where one chooses the bin width v to minimize

$$K(v, M) = \frac{1}{v}\left(\frac{2}{N-1} - \frac{N+1}{N-1}\sum_{m=1}^{M} \pi_i^2\right) \tag{4}$$

where M is the number of bins and π_i are the bin probabilities. Rudemo obtains a similar rule by applying cross-validation techniques with a Kullback-Leibler risk function [5].

We approach this problem from a different perspective. Since the underlying density is not known, it is not reasonable to use an optimization criterion that relies on the error between our density model and an unknown true density. Instead, we consider

E-mail address: kknuth@albany.edu.

https://doi.org/10.1016/j.dsp.2019.102581
1051-2004/

the histogram to be a piecewise-constant model of the underlying probability density. Using Bayesian probability theory we derive a straightforward algorithm that computes the posterior probability of the number of bins for a given data set. Within this framework defined by our likelihood and prior probability assignments, one can objectively select an optimal piecewise-constant model describing the density function from which the data were sampled.

It should be emphasized that this paper considers equal bin width piecewise-constant density models where one possesses little to no prior information about the underlying density from which the data were sampled. In many applications, variable bin width models [6–10] may be more efficient or appropriate, and certainly if one possesses prior information about the underlying density, a more appropriate model should be considered.

2. The piecewise-constant density model

We are given a dataset consisting of N data values that were sampled from an unknown probability density function. The sampled data values are assumed to be known precisely so that there is no additional measurement uncertainty associated with each datum. We begin by considering the histogram as a piecewise-constant model of the probability density function from which N data points were sampled. This model has M bins with each bin having equal width $v = v_k$, where k is used to index the bins. Together they encompass an entire range of data values $V = Mv$. Note that for a one-dimensional histogram, v_k is the width of the kth bin. In the case of a multi-dimensional histogram, this will be a multi-dimensional volume. Each bin has a "height" h_k, which is the constant probability density over the region of the bin. Integrating this constant probability density h_k over the width of the bin v_k leads to a probability mass of $\pi_k = h_k v_k$ for the bin. This results in the following piecewise-constant model $h(x)$ of the unknown probability density function $f(x)$

$$h(x) = \sum_{k=1}^{M} h_k \, \Pi(x_{k-1}, x, x_k), \tag{5}$$

where h_k is the probability density of the kth bin with edges defined by x_{k-1} and x_k, and $\Pi(x_{k-1}, x, x_k)$ is the boxcar function where

$$\Pi(x_a, x, x_b) = \begin{cases} 0 & \text{if } x < x_a \\ 1 & \text{if } x_a \le x < x_b \\ 0 & \text{if } x_b \le x \end{cases} \tag{6}$$

This density model can be re-written in terms of the bin probabilities π_k as

$$h(x) = \frac{M}{V} \sum_{k=1}^{M} \pi_k \, \Pi(x_{k-1}, x, x_k). \tag{7}$$

It is important to keep in mind that $h(x)$ is not a histogram, but rather it is a piecewise-constant probability density function. The bin heights h_k represent the probability density assigned to the kth bin, and the parameters π_k represents the probability mass of the kth bin.

Given M bins and the normalization condition that the integral of the probability density equals unity, we are left with $M-1$ bin probabilities: $\pi_1, \pi_2, \ldots, \pi_{M-1}$, each describing the probability that samples will be drawn from each of the M bins. The normalization condition requires that $\pi_M = 1 - \sum_{k=1}^{M-1} \pi_k$. For simplicity, we assume that the bin alignment is fixed so that extreme data points define the edges of the extreme bins.

2.1. The likelihood of the piecewise-constant model

The likelihood function is a probability density that when multiplied by dx describes the probability that a datum d_n is found to have a value in the infinitesimal range between some number x and $x + dx$. Since we have assumed that there is no additional measurement uncertainty associated with each datum, the likelihood that d_n will have a value between x and $x + dx$ falling within the kth bin is given the uniform probability density in the region defined by that bin

$$p(d_n|\pi_k, M, I) = h_k = \frac{\pi_k}{v_k} \tag{8}$$

where I represents our prior knowledge about the problem, which includes the range of the data and the bin alignment. For equal width bins, the likelihood density reduces to

$$p(d_n|\pi_k, M, I) = \frac{M}{V}\pi_k. \tag{9}$$

For N independently sampled data points, the joint likelihood is given by

$$p(\underline{d}|\underline{\pi}, M, I) = \left(\frac{M}{V}\right)^N \pi_1^{n_1}\pi_2^{n_2}\ldots\pi_{M-1}^{n_{M-1}}\pi_M^{n_M} \tag{10}$$

where $\underline{d} = \{d_1, d_2, \ldots, d_N\}$, $\underline{\pi} = \{\pi_1, \pi_2, \ldots, \pi_{M-1}\}$, and the n_i are the number of data points in the ith bin. Equation (10) is data-dependent and describes the likelihood that the hypothesized piecewise-constant model accounts for the data. Individuals who recognize this as having the form of the multinomial distribution may be tempted to include its familiar normalization factor. However, it is important to note that this likelihood function is properly normalized as is, which we now demonstrate. For a single datum d, the likelihood that it will take the value x is

$$p(d = x|\underline{\pi}, M, I) = \frac{1}{v}\sum_{k=1}^{M} \pi_k \, \Pi(x_{k-1}, x, x_k), \tag{11}$$

where we have written $v = \frac{V}{M}$. Multiplying the probability density by dx to get the probability and integrating over all possible values of x we have

$$\begin{aligned} \int_{-\infty}^{\infty} dx \, p(d = x|\underline{\pi}, M, I) &= \int_{-\infty}^{\infty} dx \, \frac{1}{v}\sum_{k=1}^{M} \pi_k \, \Pi(x_{k-1}, x, x_k) \\ &= \frac{1}{v}\sum_{k=1}^{M} \int_{-\infty}^{\infty} dx \, \pi_k \, \Pi(x_{k-1}, x, x_k) \\ &= \frac{1}{v}\sum_{k=1}^{M} \pi_k \, v \\ &= \sum_{k=1}^{M} \pi_k \\ &= 1. \end{aligned} \tag{12}$$

2.2. The prior probabilities

For the prior probability of the number of bins, we assign a uniform density

$$p(M|I) = \begin{cases} C^{-1} & \text{if } 1 \le M \le C \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where C is the maximum number of bins to be considered. This could reasonably be set to the range of the data divided by smallest non-zero distance between any two data points.

We assign a non-informative prior for the bin parameters $\pi_1, \pi_2, \ldots, \pi_{M-1}$, the possible values of which lie within a simplex defined by the corners of an M-dimensional hypercube with unit side lengths

$$p(\underline{\pi}|M, I) = \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M}\left[\pi_1\pi_2\cdots\pi_{M-1}\left(1-\sum_{i=1}^{M-1}\pi_i\right)\right]^{-1/2}. \quad (14)$$

Equation (14) is the Jeffreys's prior for the multinomial likelihood (10) [11–13], and has the advantage in that it is also the conjugate prior to the multinomial likelihood. The result is that the posterior probability is a Dirichlet-multinomial distribution, which is widely used in machine learning [14]. A similar posterior probability is used by Endres and Foldiak [9] to solve the more general problem of variable-width bin models.

2.3. The posterior probability

Using Bayes' Theorem, the posterior probability of the histogram model is proportional to the product of the priors and the likelihood

$$p(\underline{\pi}, M|\underline{d}, I) \propto p(\underline{\pi}|M, I)\; p(M|I)\; p(\underline{d}|\underline{\pi}, M, I). \quad (15)$$

Substituting (10), (13), and (14) gives the joint posterior probability for the piecewise-constant density model

$$p(\underline{\pi}, M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \times \pi_1^{n_1-\frac{1}{2}}\pi_2^{n_2-\frac{1}{2}}\ldots\pi_{M-1}^{n_{M-1}-\frac{1}{2}}\left(1-\sum_{i=1}^{M-1}\pi_i\right)^{n_M-\frac{1}{2}}, \quad (16)$$

where $p(M|I)$ is absorbed into the implicit proportionality constant with the understanding that we will only consider a reasonable range of bin numbers.

The goal is to obtain the posterior probability for the number of bins M. To do this we integrate the joint posterior over all possible values of $\pi_1, \pi_2, \ldots, \pi_{M-1}$ in the simplex. While the result (30) is well-known [15,14], it is instructive to see how such integrations can be handled. The expression we desire is written as a series of nested integrals over the $M-1$ dimensional parameter space of bin probabilities

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \int_0^1 d\pi_1\, \pi_1^{n_1-\frac{1}{2}} \int_0^{1-\pi_1} d\pi_2\, \pi_2^{n_2-\frac{1}{2}} \ldots \quad (17)$$

$$\ldots \int_0^{(1-\sum_{i=1}^{M-2}\pi_i)} d\pi_{M-1}\, \pi_{M-1}^{n_{M-1}-\frac{1}{2}}\left(1-\sum_{i=1}^{M-1}\pi_i\right)^{n_M-\frac{1}{2}}.$$

In order to write this more compactly, we first define

$$a_1 = 1$$
$$a_2 = 1-\pi_1$$
$$a_3 = 1-\pi_1-\pi_2$$
$$\vdots$$
$$a_{M-1} = 1-\sum_{k=1}^{M-2}\pi_k \quad (18)$$

and note the recursion relation

$$a_k = a_{k-1} - \pi_{k-1}. \quad (19)$$

These definitions greatly simplify the sum in the last term as well as the limits of integration

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \int_0^{a_1} d\pi_1\, \pi_1^{n_1-\frac{1}{2}} \int_0^{a_2} d\pi_2\, \pi_2^{n_2-\frac{1}{2}} \ldots$$
$$\ldots \int_0^{a_{M-1}} d\pi_{M-1}\, \pi_{M-1}^{n_{M-1}-\frac{1}{2}}(a_{M-1}-\pi_{M-1})^{n_M-\frac{1}{2}}. \quad (20)$$

To solve the set of nested integrals in (17), consider the general integral

$$I_k = \int_0^{a_k} d\pi_k\, \pi_k^{n_k-\frac{1}{2}}(a_k-\pi_k)^{b_k}. \quad (21)$$

This integral can be re-written as

$$I_k = a_k^{b_k}\int_0^{a_k} d\pi_k\, \pi_k^{n_k-\frac{1}{2}}\left(1-\frac{\pi_k}{a_k}\right)^{b_k}. \quad (22)$$

Setting $u = \frac{\pi_k}{a_k}$ we have

$$I_k = a_k^{b_k}\int_0^1 du\, a_k^{n_k+\frac{1}{2}}u^{n_k-\frac{1}{2}}(1-u)^{b_k}$$
$$= a_k^{b_k+n_k+\frac{1}{2}}\int_0^1 du\, u^{n_k-\frac{1}{2}}(1-u)^{b_k},$$
$$= a_k^{b_k+n_k+\frac{1}{2}}B(n_k+\frac{1}{2}, b_k+1) \quad (23)$$

where $B(\cdot)$ is the Beta function with

$$B(n_k+\frac{1}{2}, b_k+1) = \frac{\Gamma\left(n_k+\frac{1}{2}\right)\Gamma(b_k+1)}{\Gamma\left(n_k+\frac{1}{2}+b_k+1\right)}. \quad (24)$$

To solve all of the integrals we rewrite a_k in (23) using the recursion formula (19)

$$I_k = (a_{k-1}-\pi_{k-1})^{b_k+n_k+\frac{1}{2}}B(n_k+\frac{1}{2}, b_k+1). \quad (25)$$

By defining

$$b_{M-1} = n_M - \frac{1}{2}$$
$$b_{k-1} = b_k + n_k + \frac{1}{2} \quad (26)$$

we find

$$b_1 = N - n_1 + \frac{M}{2} - \frac{3}{2}. \quad (27)$$

Finally, integrating (20) gives

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \prod_{k=1}^{M-1} B(n_k + \frac{1}{2}, b_k + 1),$$

which can be simplified further by expanding the Beta functions using (24)

$$\begin{aligned} p(M|\underline{d}, I) \propto & \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \\ & \times \frac{\Gamma(n_1 + \frac{1}{2})\Gamma(b_1 + 1)}{\Gamma(n_1 + \frac{1}{2} + b_1 + 1)} \times \frac{\Gamma(n_2 + \frac{1}{2})\Gamma(b_2 + 1)}{\Gamma(n_2 + \frac{1}{2} + b_2 + 1)} \\ & \times \ldots \times \frac{\Gamma(n_{M-1} + \frac{1}{2})\Gamma(b_{M-1} + 1)}{\Gamma(n_{M-1} + \frac{1}{2} + b_{M-1} + 1)} \end{aligned} \tag{28}$$

Using the recursion relation (26) for the b_k, we see that the general term $\Gamma(b_k + 1)$ in each numerator, except the last, cancels with the denominator in the following term. This leaves

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \frac{\prod_{k=1}^{M} \Gamma(n_k + \frac{1}{2})}{\Gamma(n_1 + b_1 + \frac{3}{2})}, \tag{29}$$

where we have used (26) to observe that $\Gamma(b_{M-1} + 1) = \Gamma(n_M + 1/2)$. Last, again using the recursion relation in (26) we find that $b_1 = N - n_1 + \frac{M}{2} - \frac{3}{2}$, which results in our marginal posterior probability

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \frac{\prod_{k=1}^{M} \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}. \tag{30}$$

The implicit proportionality constant is Z^{-1}, where Z is the Bayesian evidence found by summing over all the possible numbers of bins

$$Z = \sum_{M=1}^{M_{max}} p(M|\underline{d}, I). \tag{31}$$

Since the evidence is found by summing over the numbers of bins, it depends only on the data $\underline{d}$. For this reason, we will work with the un-normalized posterior, and shall refer to its values as relative posterior probabilities.

In optimization problems, it is often easier to maximize the logarithm of the posterior

$$\begin{aligned} \log p(M|\underline{d}, I) = & N \log M + \log\Gamma\left(\frac{M}{2}\right) - M \log\Gamma\left(\frac{1}{2}\right) \\ & - \log\Gamma\left(N + \frac{M}{2}\right) + \sum_{k=1}^{M} \log\Gamma\left(n_k + \frac{1}{2}\right) \\ & + K, \end{aligned} \tag{32}$$

where K represents the sum of the volume term and the logarithm of the implicit proportionality constant, which is the inverse of the Bayesian evidence. The optimal number of bins $\hat{M}$ is found by identifying the mode of the logarithm of the marginal posterior

$$\hat{M} = \arg\max_M \{\log p(M|\underline{d}, I)\}. \tag{33}$$

Such a result is reassuring, since it is independent of the order in which the bins are counted. Many software packages are equipped to quickly compute the log of the gamma function. However, for more basic implementations, the following definitions from Abramowitz and Stegun [16] can be used for integer m:

$$\log\Gamma(m) = \sum_{k=1}^{m-1} \log k \tag{34}$$

$$\log\Gamma\left(m + \frac{1}{2}\right) = \frac{1}{2}\log\pi - m\log 2 + \sum_{k=1}^{m} \log(2k - 1). \tag{35}$$

Equation (32) allows one to easily identify the number of bins M which optimize the posterior. We call this the `OPTBINS` algorithm[1] and provide the `Matlab` code in the Appendix.

3. The posterior probability for the bin height

In order to obtain the posterior probability for the probability mass of a particular bin, we begin with the joint posterior (16) and integrate over all the other bin probability masses. Since we can consider the bins in any order, the resulting expression is similar to the multiple nested integral in (17) except that the integral for one of the $M - 1$ bins is not performed. Treating the number of bins as a given, we can use the product rule to get

$$p(\underline{\pi}|\underline{d}, M, I) = \frac{p(\underline{\pi}, M|\underline{d}, I)}{p(M|\underline{d}, I)} \tag{36}$$

where the numerator is given by (16) and the denominator by (30). Since the bins can be treated in any order, we derive the marginal posterior for the first bin and generalize the result for the kth bin. The marginal posterior is

$$\begin{aligned} p(\pi_1|\underline{d}, M, I) = & \frac{(\frac{M}{V})^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M}}{p(M|\underline{d}, I)} \pi_1^{n_1 - \frac{1}{2}} \\ & \times \int_0^{a_2} d\pi_2\, \pi_2^{n_2 - \frac{1}{2}} \int_0^{a_3} d\pi_3\, \pi_3^{n_3 - \frac{1}{2}} \ldots \\ & \ldots \int_0^{a_{M-1}} d\pi_{M-1}\, \pi_{M-1}^{n_{M-1} - \frac{1}{2}} (a_{M-1} - \pi_{M-1})^{n_M - \frac{1}{2}}. \end{aligned} \tag{37}$$

Evaluating the integrals and substituting (28) into the denominator we get

$$p(\pi_1|\underline{d}, M, I) = \frac{\prod_{k=2}^{M-1} B(n_k + \frac{1}{2}, b_k + 1)}{\prod_{k=1}^{M-1} B(n_k + \frac{1}{2}, b_k + 1)} \pi_1^{n_1 - \frac{1}{2}} (1 - \pi_1)^{b_1}. \tag{38}$$

Canceling terms and explicitly writing b_1, the marginal posterior for π_1 is

$$\begin{aligned} & p(\pi_1|\underline{d}, M, I) \\ & = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_1 + \frac{1}{2})\Gamma(N - n_1 + \frac{M-1}{2})} \pi_1^{n_1 - \frac{1}{2}} (1 - \pi_1)^{N - n_1 + \frac{M-3}{2}}, \end{aligned} \tag{39}$$

which can easily be verified to be normalized by integrating π_1 over its entire possible range from 0 to 1. Since the bins can be considered in any order, this is a general result for the kth bin

$$\begin{aligned} & p(\pi_k|\underline{d}, M, I) \\ & = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \pi_k^{n_k - \frac{1}{2}} (1 - \pi_k)^{N - n_k + \frac{M-3}{2}}. \end{aligned} \tag{40}$$

[1] The method has also been referred to as the Knuth method.

The mean bin probability mass can be found from its expectation

$$\langle \pi_k \rangle = \int_0^1 d\pi_k \, \pi_k \, p(\pi_k | \underline{d}, M, I), \tag{41}$$

which substituting (40) gives

$$\langle \pi_k \rangle = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \times \int_0^1 d\pi_k \, \pi_k^{n_k + \frac{1}{2}} (1 - \pi_k)^{N - n_k + \frac{M-3}{2}}. \tag{42}$$

The integral again gives a Beta function, which when written in terms of Gamma functions is

$$\langle \pi_k \rangle = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \times \frac{\Gamma(n_k + \frac{3}{2})\Gamma(N - n_k + \frac{M-1}{2})}{\Gamma(N + \frac{M}{2} + 1)}. \tag{43}$$

Using the fact that $\Gamma(x+1) = x\Gamma(x)$ and canceling like terms, we find that

$$\langle \pi_k \rangle = \frac{n_k + \frac{1}{2}}{N + \frac{M}{2}}. \tag{44}$$

The mean probability density for bin k (the bin height) is simply

$$\mu_k = \langle h_k \rangle = \frac{\langle \pi_k \rangle}{v_k} = \left(\frac{M}{V}\right)\left(\frac{n_k + \frac{1}{2}}{N + \frac{M}{2}}\right). \tag{45}$$

It is an interesting result that bins with no counts still have a non-zero probability. This makes sense since no lack of evidence can ever prove conclusively that an event occurring in a given bin is impossible—just less probable. The Jeffrey's prior effectively places one-half of a datum in each bin.

The variance of the probability mass of the kth bin is found similarly by

$$\sigma_k^2 = \left(\frac{M}{V}\right)^2 \left(\langle \pi_k^2 \rangle - \langle \pi_k \rangle^2\right), \tag{46}$$

which gives

$$\sigma_k^2 = \left(\frac{M}{V}\right)^2 \left(\frac{(n_k + \frac{1}{2})(N - n_k + \frac{M-1}{2})}{(N + \frac{M}{2} + 1)(N + \frac{M}{2})^2}\right). \tag{47}$$

Thus, given the optimal number of bins found by maximizing (32), the mean and variance of the bin probabilities are found from (45) and (47), which allow us to construct an explicit histogram model of the probability density and perform computations replete with proper error analysis. Note that in the case where there is one bin (47) gives a zero variance.

4. Results

4.1. Demonstration using one-dimensional histograms

In this section we demonstrate the utility of this method for determining the optimal number of bins in a piecewise-constant density model by applying this method to several different data sets. Note that since it is computationally costly to marginalize the posterior probability (30) to obtain the appropriate normalization factor, the analyses below rely on the un-normalized posterior probability, the logarithms of which will be referred to the *relative log posterior*. We consider four different test cases where we have sampled 1000 data points from each of the four different probability density functions.

The first example considers a Gaussian probability density $\mathcal{N}(0,1)$. The optimal piecewise-constant density model for the 1000 data points sample d from this distribution is shown in Fig. 1A, where it is superimposed over a 100-bin histogram that better illustrates the locations of the sampled points. Fig. 1B shows that the relative log posterior probability (32) peaks at 14 bins. Note that the bin heights for the piecewise-constant density model are determined from (44); whereas the bin heights of the 100-bin histogram illustrating the data samples are proportional to the counts. For this reason, the two pictures are not directly comparable.

The second example considers a 4-step constant piecewise density. Fig. 1C shows the optimal binning for the 1000 sampled data points. The relative log posterior (Fig. 1D) peaks at 4 bins, which indicates that the method correctly detects the 4-step structure.

A uniform density is used to sample 1000 data points in the third example. Figs. 1E and 1F, demonstrate that samples drawn from a uniform density were best described by a single bin. This result is significant, since entropy estimates computed from these data would be biased if multiple bins were used to describe the distribution of the sampled data.

Last, we consider a density function that consists of a mixture of three sharply-peaked Gaussians with a uniform background (Fig. 1G). The posterior peaks at 52 bins indicating that the data warrant a detailed model (Fig. 1H). The spikes in the relative log posterior are due to the fact that the bin edges are fixed. The relative log posterior is large at values of M where the bins happen to line up with the Gaussians, and small when they are misaligned. This last example demonstrates one of the weaknesses of the equal bin-width model, as many bins are needed to describe the uniform density between the three narrow peaks. In addition, the lack of an obvious peak indicates that there is a range of bin numbers that will result in reasonable models.

5. Effects of small sample size

5.1. Small samples and asymptotic behavior

It is instructive to observe how this algorithm behaves in situations involving small sample sizes. We begin by considering the extreme case of two data points $N = 2$. In the case of a single bin, $M = 1$, the posterior probability reduces to

$$p(M=1|d_1,d_2,I) \propto M^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})} \propto 1^2 \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2})^1} \frac{\Gamma(2+\frac{1}{2})}{\Gamma(2+\frac{1}{2})} = 1, \tag{48}$$

so that the log posterior is zero. For $M > 1$, the two data points lie in separate bins, resulting in

$$\begin{aligned} p(M|d_1,d_2,I) &\propto M^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})} \\ &\propto M^2 \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\Gamma(1+\frac{1}{2})^2 \Gamma(\frac{1}{2})^{M-2}}{\Gamma(2+\frac{M}{2})} \\ &\propto M^2 \frac{\Gamma(\frac{3}{2})^2}{\Gamma(\frac{1}{2})^2} \frac{\Gamma(\frac{M}{2})}{\Gamma(2+\frac{M}{2})} \\ &\propto \frac{1}{2} \cdot \frac{M}{1+\frac{M}{2}}. \end{aligned} \tag{49}$$

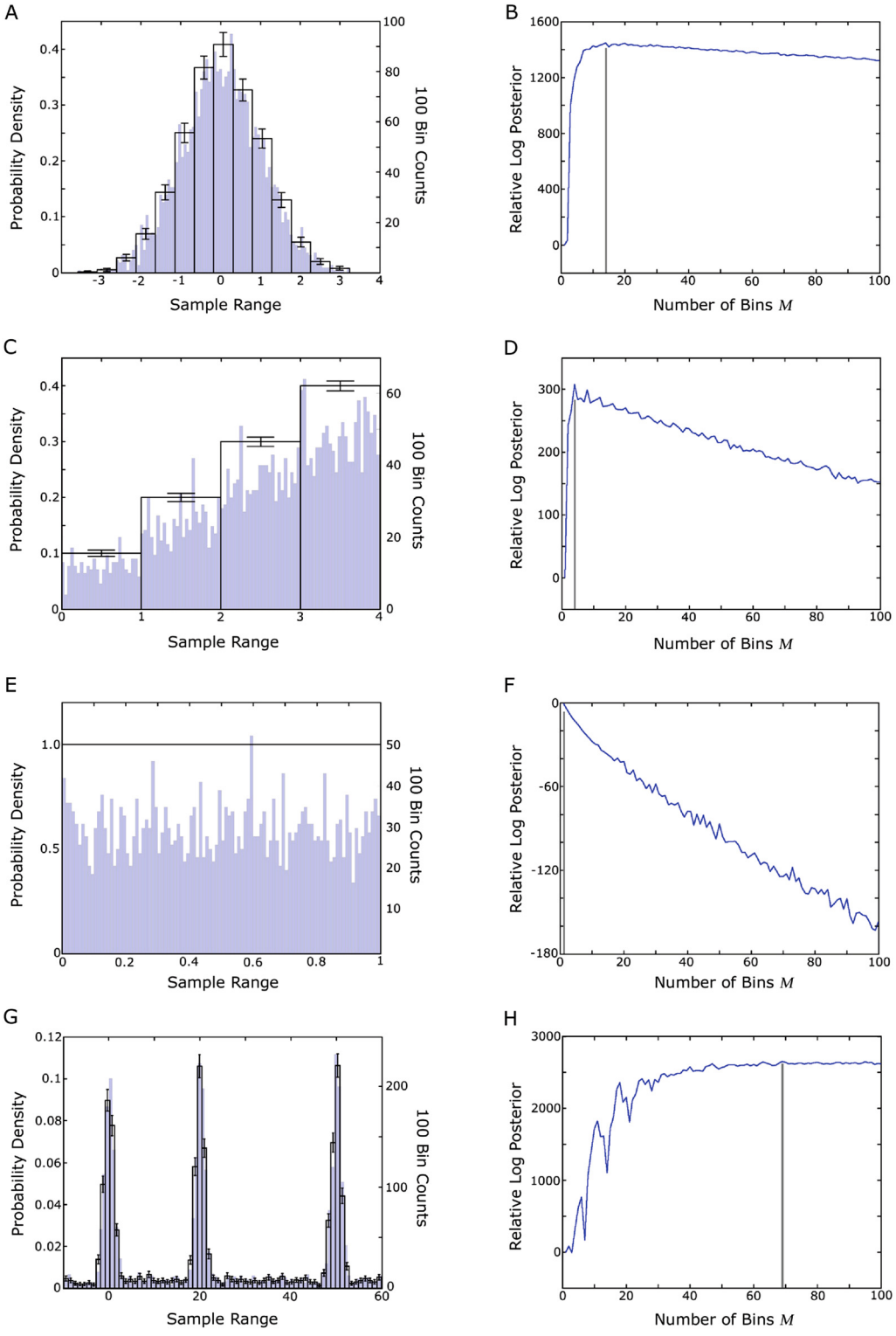

Fig. 1. To demonstrate the technique, 1000 samples were sampled from four different probability density functions. (A) The optimal piecewise-constant model for 1000 samples drawn from a Gaussian density function is superimposed over a 100-bin histogram that shows the distribution of data samples. (B) The relative log posterior probability of the number of bins peaks at 14 bins for these 1000 data sampled from the Gaussian density. (C) Samples shown are from a 4-step piecewise-constant density function. The relative log posterior peaks at four bins (D) indicating that the method correctly detects the four-step structure. (E) These data were sampled from a uniform density as verified by the relative log posterior probability shown in (F), which starts at a maximum value of one and decreases with increasing numbers of bins. (G) Here we demonstrate a more complex example—three Gaussian peaks plus a uniform background. (H) The posterior, which peaks at 52 bins, demonstrates clearly that the data themselves support this detailed picture of the pdf.

Fig. 2A shows the log posterior which starts at zero for a single bin, drops to $\log(\frac{1}{2})$ for $M = 2$ and then increases monotonically approaching zero in the limit as M goes to infinity. The result is that a single bin is the most probable solution for two data points.

For three data points in a single bin ($N = 3$ and $M = 1$), the posterior probability is one, resulting in a log posterior of zero. In the $M > 1$ case where there are two data points in one bin and one datum point in another, the posterior probability is

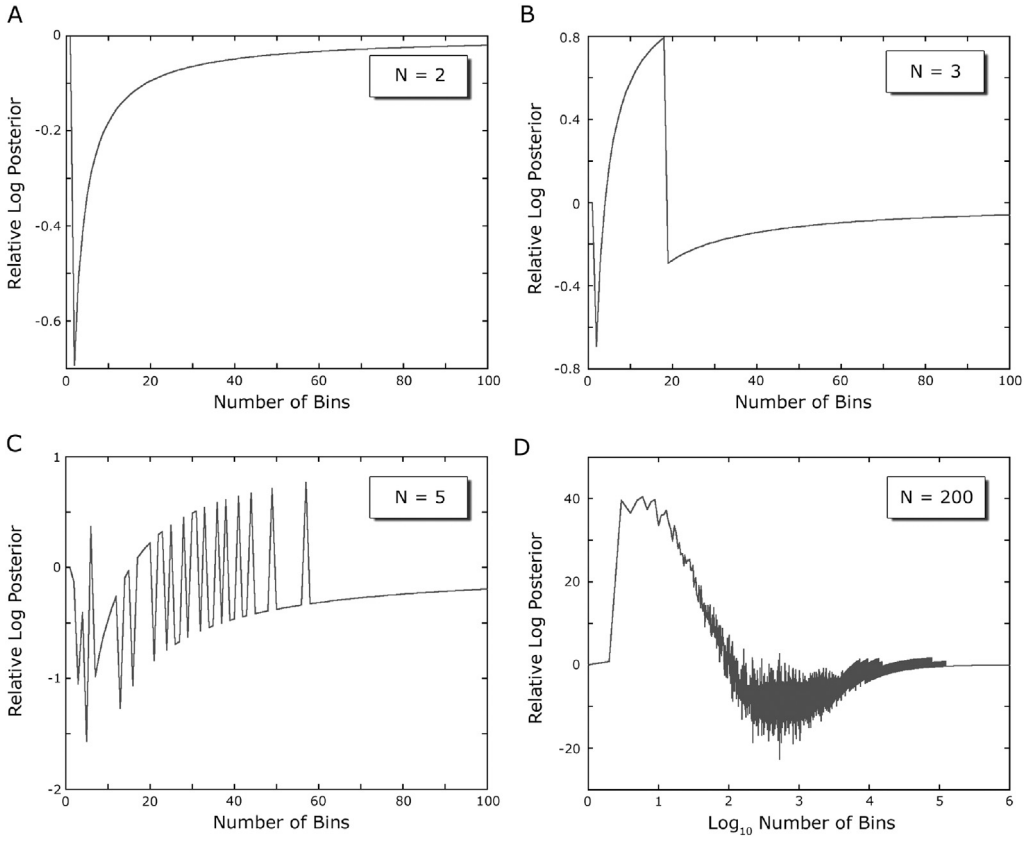

Fig. 2. These figures demonstrate the behavior of the relative log posterior for small numbers of samples. (A) With only $N = 2$ samples, the log posterior is maximum when the samples are in the same bin $M = 1$. For $M > 1$, the log posterior follows the function described in (49) in the text. (B) The relative log posterior is slightly more complicated for $N = 3$. For $M = 1$ all three points lie in the same bin. As M increases, two data points are in one bin and the remaining datum point is in another bin. The functional form is described by (50). Eventually, all three data points lie in separate bins and the relative log posterior is given by (51). (C) The situation is more complicated still for $N = 5$ data points. As M increases, a point is reached when, depending on the particular value of M, the points will be in separate bins. As M changes value, two points may again fall into the same bin. This gives rise to this oscillation in the log posterior. Once all points are in separate bins, the behavior follows a well-defined functional form (52). (D) This plot shows the behavior for a large number of data points $N = 200$. The log posterior now displays a more well-defined mode indicating that there is a well-defined optimal number of bins. As M approaches 10000 to 100000 bins, one can see some of the oscillatory behavior demonstrated in the small N cases.

$$p(M|d_1, d_2, d_3, I) \propto \frac{3}{4} \cdot \frac{M^2}{(2 + \frac{M}{2})(1 + \frac{M}{2})}, \tag{50}$$

and for each point in a separate bin we have

$$p(M|d_1, d_2, d_3, I) \propto \frac{1}{4} \cdot \frac{M^2}{(2 + \frac{M}{2})(1 + \frac{M}{2})}. \tag{51}$$

While the logarithm of the un-normalized posterior in (50) can be greater than zero, as M increases, the data points eventually fall into separate bins. This causes the posterior to change from (50) to (51) resulting in a dramatic decrease in the logarithm of the posterior, which then asymptotically increases to zero as $M \to \infty$. This behavior is shown in Fig. 2B.

More rich behavior can be seen in the case of $N = 5$ data points. The results again (Fig. 2C) depend on the relative positions of the data points with respect to one another. In this case the posterior probability switches between two types of behavior as the number of bins increase depending on whether the bin positions force two data points together in the same bin or separate them into two bins. The ultimate result is a ridiculous *maximum a posteriori* solution of 57 bins. Clearly, for a small number of data points, the mode depends sensitively on the relative positions of the samples in a way that is not meaningful. In these cases there are too few data points to model a density function.

With a larger number of samples, the posterior probability shows a well-defined mode indicating a well-determined optimal number of bins. In the general case of $M > N$ where each of the N data points is in a separate bin, we have

$$p(M|\underline{d}, I) \propto \left(\frac{M}{2}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(N + \frac{M}{2}\right)}, \tag{52}$$

which again results in a log posterior that asymptotically approaches zero as $M \to \infty$. Fig. 2D demonstrates these two effects for $N = 200$. This also can be compared to the log posterior for 1000 Gaussian samples in Fig. 1B.

5.2. Sufficient data

The investigation on the effects of small sample size in the previous section raises the question as to how many data points are needed to estimate the probability density function. The general shape of a healthy log posterior reflects a sharp initial rise to a well-defined peak, and a gradual fall-off as the number of bins M increases from one (e.g. Fig. 1B, Fig. 2D). With small sample sizes, however, one finds that the bin heights have large error bars (Fig. 3A) so that $\mu_i \simeq \sigma_i$, and that the log posterior is multi-modal (Fig. 3B) with no clear peak.

We tested our algorithm on data sets with 199 different sample sizes from $N = 2$ to $N = 200$. One thousand data sets were drawn from a Gaussian distribution for each value of N. The standard deviation of the number of bins obtained for these 1000 data sets at a given value if N was used as an indicator of the stability of the solution.

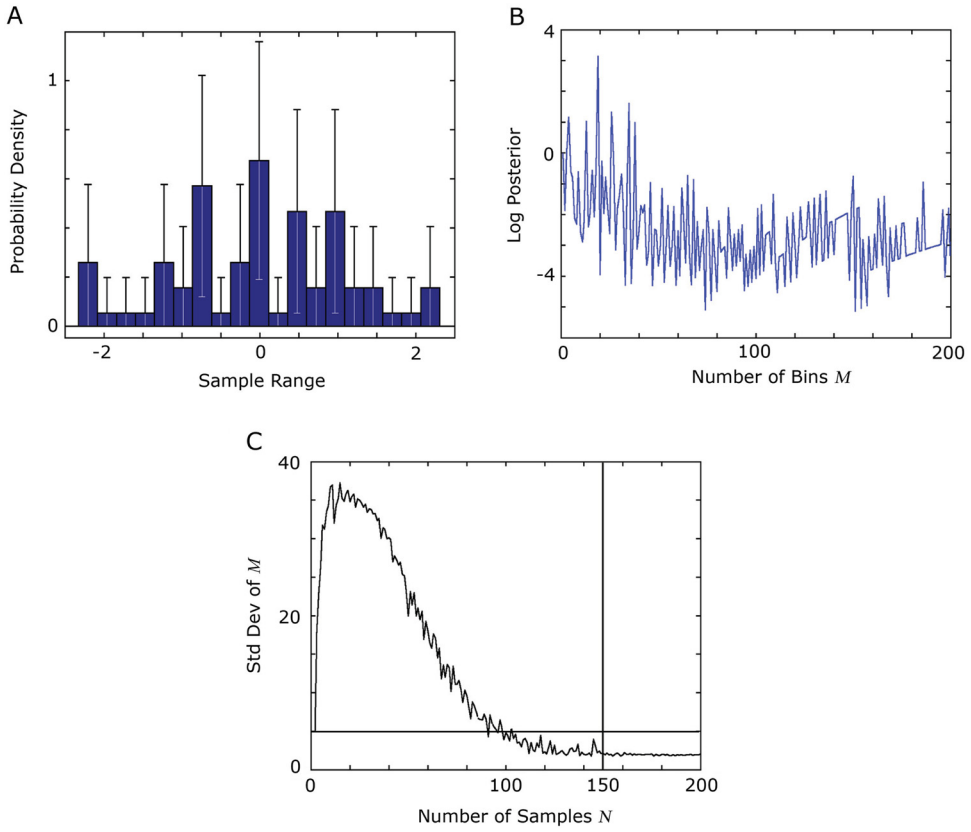

Fig. 3. (A) An optimal density model ($M = 19$) for $N = 30$ data points sampled from a Gaussian distribution. The fact that the error bars on the bin probabilities are as large as the probabilities themselves indicates that this is a poor estimate. (B) The log posterior probability for the number of bins possesses no well-defined peak, and is instead reminiscent of noise. (C) This plot shows the standard deviation of the estimated number of bins M for 1000 data sets of N points, ranging from 2 to 200, sampled from a Gaussian distribution. The standard deviation stabilizes around $\sigma_M = 2$ bins for $N > 150$ indicating the inherent level of uncertainty in the problem. This suggests that one requires at least 150 data points to consistently perform such probability density estimates, and can perhaps get by with as few as 100 data points in some cases.

Fig. 3C shows a plot of the standard deviation of the number of bins selected for the 1000 data sets at each value of N. As we found above, with two data points, the optimal solution is always one bin giving a standard deviation of zero. This increases dramatically as the number of data points increases, as we saw in our example with $N = 5$ and $M = 57$. This peaks around $N = 15$ and slowly decreases as N increases further. The standard deviation of the number of bins decreased to $\sigma_M < 5$ for $N > 100$, and stabilized to $\sigma_M \simeq 2$ for $N > 150$.

While 30 samples may be sufficient for estimating the mean and variance of a density function known to be Gaussian, it is clear that more samples are needed to reliably estimate the shape of an unknown density function. In the case where the data are described by a Gaussian, it would appear that at least 150 samples, are required to accurately and consistently infer the shape of a one-dimensional density function. By examining the shape of the log posterior, one can easily determine whether one has sufficient data to estimate the density function. In the event that there are too few samples to perform such estimates, one can either incorporate additional prior information or collect more data.

6. Digitized data

Due to the way that computers represent data, all data are essentially represented by integers [17]. In some cases, the data samples have been intentionally rounded or truncated, often to save storage space or transmission time. It is well-known that any non-invertible transformation, such as rounding, destroys information. Here we investigate how severe losses of information due to rounding or truncation affects the `OPTBINS` algorithm.

When data are digitized via truncation or rounding, the digitization is performed so as to maintain a resolution that we will denote by Δx. That is, if the data set has values that range from 0 to 1, and we represent these numbers with eight bits, the minimum resolution we can maintain is $\Delta x = 1/2^8 = 1/256$. For a sufficiently large data set (in this example $N > 256$) the pigeonhole principle indicates that it will be impossible to have a situation where each datum is in its own bin when the number of bins is greater than a critical number, $M > M_{\Delta x}$, where

$$M_{\Delta x} = \frac{V}{\Delta x}, \tag{53}$$

and V is the range of the data considered (see Fig. 4). Once $M > M_{\Delta x}$ the number of populated bins P will remain unchanged since the bin width w for $M > M_{\Delta x}$ will be smaller then the digitization resolution, $w < \Delta x$.

For all bin numbers $M > M_{\Delta x}$, there will be P populated bins with populations $n_1, n_2, \ldots, n_P$. This leads to a form for the marginal posterior probability for M (30) that depends only on the number of instances of each discrete value that was recorded, $n_1, n_2, \ldots, n_P$. Since these values do not vary for $M > M_{\Delta x}$, the marginal posterior can be expressed solely as a function of M

$$p(M|\underline{d}, I) \propto \left(\frac{M}{2}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(N + \frac{M}{2}\right)} \cdot 2^N \frac{\prod_{p=1}^{P} \Gamma(n_p + \frac{1}{2})}{\Gamma\left(\frac{1}{2}\right)^P}, \tag{54}$$

where the product over p is over populated bins only. Comparing this to (52), the function on the right-hand side asymptotically approaches a value greater than one so that its logarithm increases asymptotically to a value greater than zero.

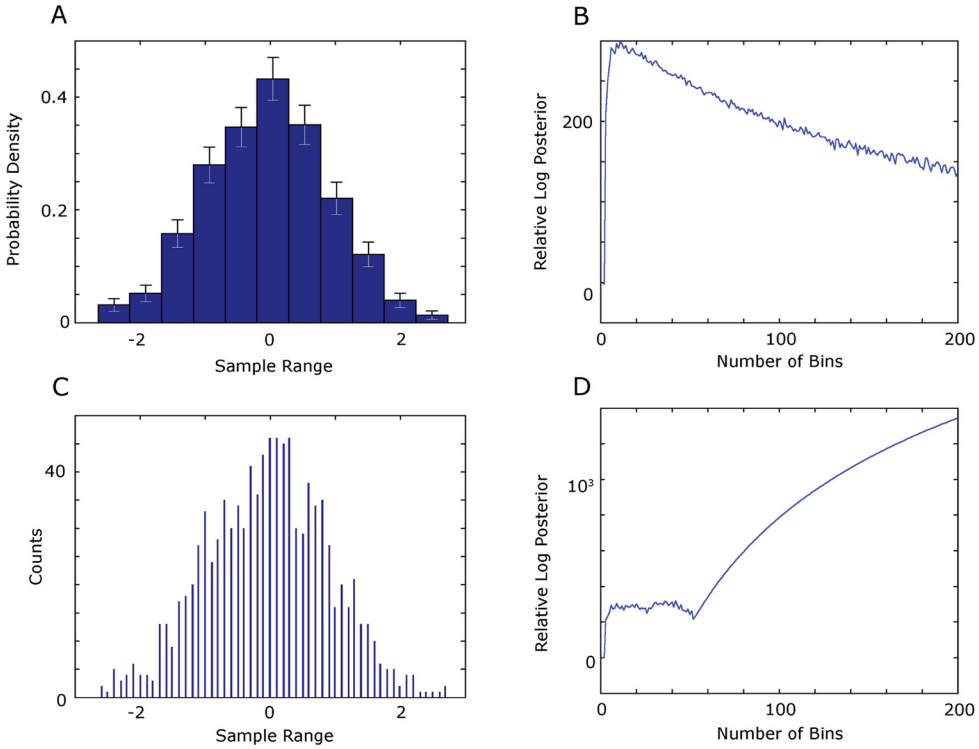

Fig. 4. $N = 1000$ data points were sampled from a Gaussian distribution $\mathcal{N}(0, 1)$. The top plots show (A) the estimated density function using optimal binning and (B) the relative log posterior, which exhibits a well-defined peak at $M = 11$ bins. The bottom plots reflect the results using the same data set after it has been rounded with $\Delta x = 0.1$ to keep only the first decimal place. (C) There is no optimal binning as the algorithm identifies the discrete structure as being a more salient feature than the overall Gaussian shape of the density function. (D) The relative log posterior displays no well-defined peak, and in addition, for large numbers of M displays a monotonically increasing curve given by (52) that asymptotes to a positive value. This indicates that the data have been severely rounded.

As the number of bins M increases, the point is reached where the data can not be further separated; call this point M_{crit}. In this situation, there are n_p data points in the pth bin and the posterior probability can be written as

$$p(M|\underline{d}, I) \propto \left(\frac{M}{2}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(N + \frac{M}{2}\right)} \cdot \prod_{p=1}^{P} (2n_p - 1)!!, \tag{55}$$

where !! denotes the double factorial [18]. For $M > M_{crit}$, as $M \to \infty$, the log posterior asymptotes to $\sum_{p=1}^{P} \log((2n_p - 1)!!)$, which can be further simplified to

$$\sum_{p=1}^{P} \log((2n_p - 1)!!) = (P - N)\log(2) + \sum_{p=1}^{P} \sum_{s=n_p}^{2n_p - 1} \log s. \tag{56}$$

This means that excessive truncation or rounding can be detected by comparing the mode of $\log p(M|\underline{d}, I)$ for $M < M_{crit}$ to (56) above. If the latter is larger, this indicates that the discrete nature of the data is a more significant feature than the general shape of the underlying probability density function (Fig. 4). When this is the case, a reasonable histogram model of the density function can still be obtained by adding a uniformly-distributed random number, with a range defined by the resolution Δx, to each datum point [17]. While this will produce the best histogram possible given the data, this will not recover the lost information.

7. Application to real data

It is important to evaluate the performance of an algorithm using real data. However, the greatest difficulty that such an evaluation poses is that the correct solution is unknown at best, and poorly-defined at worst. As a result, we must rely on our expectations. In these three examples, we will examine the optimal solutions obtained using `OPTBINS` and compare them to the density models obtained with both fewer and greater numbers of bins. It is expected that the histograms with fewer number of bins will be missing some essential characteristics of the density function, while the histograms with a greater number of bins will exhibit random fluctuations that appear to be unwarranted. For more definitive results, the reader is directed to Section 4.1.

In Fig. 5A we show the results from a data set called 'Abalone Data' retrieved from the UCI Machine Learning Repository [22]. The data consists of abalone weights in grams from 4177 individuals [19]. The relative log posterior (left) shows a flat plateau that has maximum at $M = 14$ bins and slowly decreases thereafter. Given this relatively flat plateau, we would expect most bin numbers in this region to produce reasonable histogram models. Compared to $M = 10$ bins (left) and $M = 20$ bins (right), the optimal density model with $M = 14$ bins captures the shape of the density function best without exhibiting what appear to be irrelevant details due to random sampling fluctuations.

Fig. 5B shows a second data set from the UCI Machine Learning Repository [22] titled 'Determinants of Plasma Retinol and Beta-Carotene Levels'. This data set provides blood plasma retinol concentrations (in ng/ml) measured from 315 individuals [20]. The optimal number of bins for this data set was determined to be $M = 9$. Although a second peak appears in the relative log posterior near $M = 15$, the rightmost density with 15 bins exhibits small random fluctuations suggesting that $M = 9$ is a better model.

Last, the Old Faithful data set is examined, which consists of 222 measurements of inter-eruption intervals rounded to the nearest minute [21]. This data is used extensively on the world wide web as an example of the difficulties in choosing bin sizes for histograms. There exists, in fact, a java applet developed by R. Webster West that allows one to interactively vary the bin size and observe the results in real time [23]. In Fig. 5D we plot the

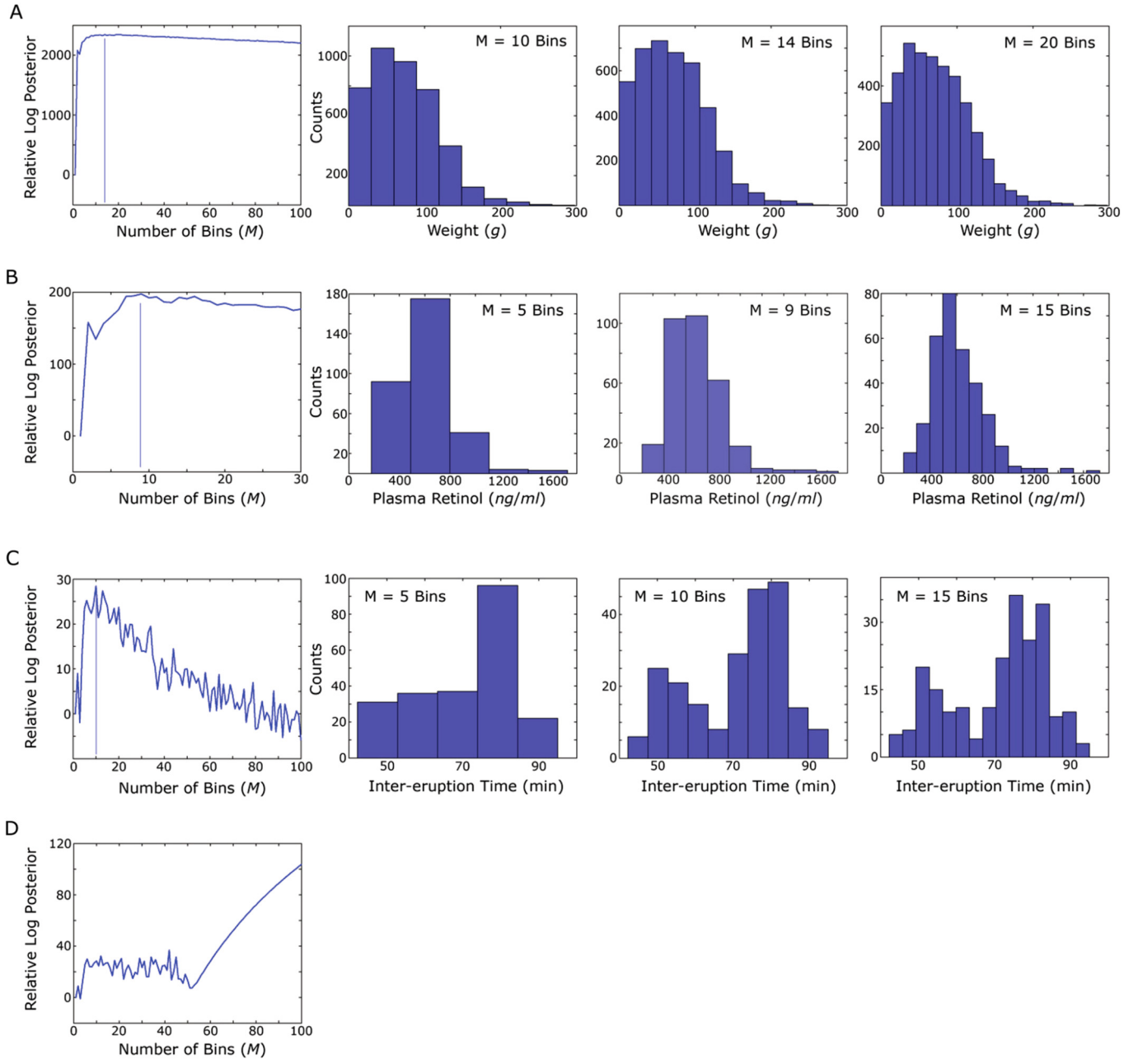

Fig. 5. The OPTBINS algorithm is used to select the number of bins for three real data sets. The relative log posterior probability of each of the three data sets is displayed on the left where the optimal number of bins is the point at which the global maximum occurs (see drop lines). The center density function in each row represents the optimally-binned model. The leftmost density function has too few bins and is not sufficiently well-resolved, whereas the rightmost density function has too many bins and is beginning to highlight irrelevant features. (A) Abalone weights in grams from 4177 individuals [19]. (B) Blood plasma retinol concentration in ng/ml measured from 315 individuals [20]. (C) The Old Faithful data set consisting of measurements of inter-eruption intervals rounded to the nearest minute [21]. In this data set, we added a uniformly distributed random number from −0.5 to 0.5 (see text). (D) This relative log posterior represents the Old Faithful inter-eruption intervals recorded to the nearest minute. Notice that the discrete nature of the data is a dominant feature as predicted by the results in the previous section. In this example, OPTBINS does more than choose the optimal number of bins, it provides a warning that the data has been severely rounded or truncated and that information has been lost.

relative log posterior for this data set. For large numbers of bins, the relative log posterior increases according to (55) as described in Section 6. This indicates that the discrete nature of the data (measured at a time resolution in minutes) is a more salient feature than the overall shape of the density function. One could have gathered more information by more carefully measuring the eruption times on the order of seconds or perhaps tens of seconds.

It is not clear whether this missing information could affect the results of previous studies, but we can obtain a useful density function by adding a small uniformly distributed number to each sample [17] as discussed in the previous section. Since the resolution is in minutes, we add a number ranging from −0.5 to 0.5 minutes. The result in Fig. 5C is a relative log posterior that has a clear maximum at $M = 10$ bins. A comparison to the $M = 5$ bin case and the $M = 20$ bin case again demonstrates that the number of bins chosen by OPTBINS results in a density function that captures the essential details and neglects the irrelevant details. In this case, our method provides additional valuable information about the data set by indicating that the discrete nature of the data was more relevant the underlying density function. This implies that a sampling strategy involving higher temporal resolution would have provided more information about the inter-eruption intervals.

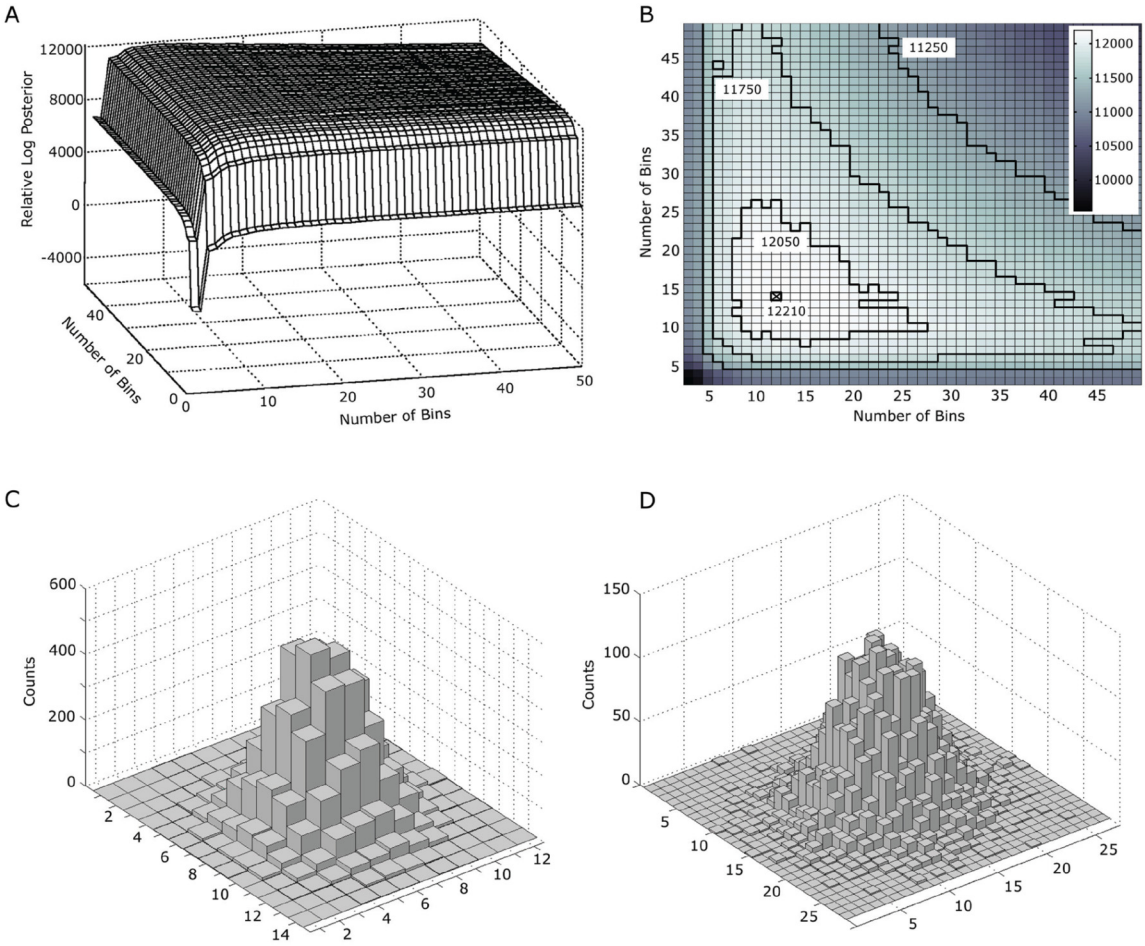

Fig. 6. 10000 samples were drawn from a two-dimensional Gaussian density to demonstrate the optimization of a two-dimensional histogram. (A) The relative logarithm of the posterior probability is plotted as a function of the number of bins in each dimension. The normalization constant has been neglected in this plot, resulting in positive values of the log posterior. (B) This plot shows the relative log posterior as a contour plot. The optimal number of bins is found to be 12×14. (C) The optimal histogram for this data set. (D) The histogram determined using Stone's method has 27×28 bins. This histogram is clearly sub-optimal since it highlights random variations that are not representative of the density function from which the data were sampled.

8. Multi-dimensional histograms

In this section, we demonstrate that our method can be extended naturally to multi-dimensional histograms. We begin by describing the method for a two-dimensional histogram. The constant-piecewise model $h(x, y)$ of the two-dimensional density function $f(x, y)$ is

$$h(x, y; M_x, M_y) = \frac{M}{V} \sum_{j=1}^{M_x} \sum_{k=1}^{M_y} \pi_{j,k} \, \Pi(x_{j-1}, x, x_j) \Pi(y_{k-1}, y, y_k), \tag{57}$$

where $M = M_x M_y$, V is the total area of the histogram, j indexes the bin labels along x, and k indexes them along y. Since the $\pi_{j,k}$ all sum to unity, we have $M - 1$ bin probability density parameters as before, where M is the total number of bins. The likelihood of obtaining a datum point d_n from bin (j, k) is still simply

$$p(d_n | \pi_{j,k}, M_x, M_y, I) = \frac{M}{V} \pi_{j,k}. \tag{58}$$

The previous prior assignments result in the posterior probability

$$p(\underline{\pi}, M_x, M_y | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \prod_{j=1}^{M_x} \prod_{k=1}^{M_y} \pi_{j,k}^{n_{j,k} - \frac{1}{2}}, \tag{59}$$

where π_{M_x, M_y} is 1 minus the sum of all the other bin probabilities. The order of the bins in the marginalization does not matter, which gives a result similar in form to the one-dimensional case

$$p(M_x, M_y | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \frac{\prod_{j=1}^{M_x} \prod_{k=1}^{M_y} \Gamma(n_{j,k} + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \tag{60}$$

where $M = M_x M_y$.

For a D-dimensional histogram, the general result is

$$p(M_1, \cdots, M_D | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \frac{\prod_{i_1=1}^{M_1} \cdots \prod_{i_D=1}^{M_D} \Gamma(n_{i_1, \ldots, i_D} + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \tag{61}$$

where M_i is the number of bins along the ith dimension, M is the total number of bins, V is the D-dimensional volume of the histogram, and $n_{i_1, \ldots, i_D}$ indicates the number of counts in the bin indexed by the coordinates $(i_1, \ldots, i_D)$. Note that the result in (32) can be used directly for a multi-dimensional histogram simply by relabeling the multi-dimensional bins with a single index.

Fig. 6 demonstrates the procedure on a data set sampled from a two-dimensional Gaussian. In this example, 10000 samples were drawn from a two-dimensional Gaussian density. Fig. 6A shows the relative logarithm of the posterior probability plotted as a function of the number of bins in each dimension. The same surface

Table 1
Summary of the histogram binning performance for $N = 500$ for data sampled from distributions consisting of the specified Number of Bins. For each algorithm, the fraction of the correct number of bins (cor) is listed as well as the mean (mean) number of bins obtained for the 100 trials. At the bottom, the overall fraction of correct number of bins (cor) is listed along with the root mean square error (rms) in the number of bin estimates for all 10000 trials. The best scores are highlighted in boldface font.

Number of bins	OPTBINS		AIC		BIC		Scott		Stone	
	cor	mean	cor	mean	cor	mean	cor	mean	cor	mean
1	**1**	**1**	0.58	2.8	0.99	**1**	0	26.8	0.59	2.7
2	**1**	**2**	0.62	3.9	0.78	1.9	0	17.2	0.64	4.6
3	**1**	**3**	0.71	4.8	0.85	**3**	0	10.9	0.66	4.9
4	**1**	**4**	0.77	6.1	0.82	3.9	0.02	7.5	0.76	6.2
5	**1**	**5**	0.7	8.3	0.86	4.8	0.27	6.2	0.68	8.3
6	0.25	5.3	0.71	8.6	**0.81**	**5.9**	0.14	4.7	0.75	8.7
7	0.76	6.5	0.67	11	**0.85**	**6.7**	0.03	3.7	0.68	10.5
8	0.29	**7.3**	0.7	11.1	**0.77**	**7.3**	0.01	3.3	0.75	10.4
9	**0.84**	**8.7**	0.8	10.8	0.77	8.4	0	2.7	0.7	12
10	**1**	**10**	0.77	13.1	0.72	9.2	0	2.4	0.74	12.9
20	**0.93**	**19.9**	0.78	23.2	0.6	15.7	0	1	0.78	24.2
25	0.59	**24.6**	**0.7**	28.6	0.43	16.4	0	0.9	0.67	29
30	0.62	**29.6**	**0.75**	32.9	0.32	16.1	0	0.2	0.73	32.1
40	**0.85**	**39.6**	0.65	43.3	0.05	8.9	0	0	0.63	43.2
50	0.58	47.9	**0.62**	50.8	0	5.7	0	0	0.57	**50.6**
60	0.07	57.2	**0.55**	**59.9**	0	4.2	0	0	0.53	58.4
70	**0.49**	**65.9**	0.45	65.4	0	3.3	0	0	0.36	61
80	0.08	**76.3**	**0.38**	71.4	0	1.9	0	0	0.33	67.1
90	**0.52**	**89.5**	0.38	77.8	0	2	0	0	0.35	71.8
100	**0.55**	**98.7**	0.24	76.3	0	1.6	0	0	0.16	68.7
	cor	rms	cor	rms	cor	rms	cor	rms	cor	rms
overall	0.48	**2.34**	**0.59**	14.52	0.21	54.98	0.005	58.17	0.54	18.07

is displayed as contour plot in Fig. 6B, where we find the optimal number of bins to be 12×14. Fig. 6C shows the optimal two-dimensional histogram model. Note that the modeled density function is displayed in terms of the number of counts rather than the probability density, which can be easily computed using (45) with error bars computed using (47). In Fig. 6D, we show the histogram obtained using Stone's method, which results in an array of 27×28 bins. This model consists of approximately four times as many bins, and as a result, random sampling variations become visible.

9. Comparison to other techniques

We now compare the OPTBINS method, also known as the Knuth method, with several other popular methods, such as the Akaike Information Criterion (AIC) [24], the Bayesian Information Criterion (BIC) [25,26], Scott's Rule (3) [1,2], and Stone's Rule (4) [4]. Since Freedman and Diaconis' (F&D) method [3] has the same functional form as Scott's Rule (3), the results using F&D are not presented here. The AIC method, applied to histograms in [27], balances the logarithm of the likelihood of the model against the number of model parameters. The number of bins is chosen to maximize

$$AIC(M) = 2 \log p(d_n|\pi_k, M, I) - 2M, \tag{62}$$

where $\log p(d_n|\pi_k, M, I)$ is the logarithm of the likelihood (10). The BIC method [26] results in a similar weighting

$$BIC(M) = 2 \log p(d_n|\pi_k, M, I) - M \log(N), \tag{63}$$

which is also maximized.

To objectively evaluate the performance of each of these algorithms, they need to be tested against data sets sampled from distributions for which the correct number of bins is known. For this reason, testing is not only limited to synthetic data, but also to data that was sampled from a probability density that is piecewise uniform with equal width bins. We considered such probability densities with the number of bins ranging from $M = 1$ to 100. The probability of each bin of the trial probability density function was selected to be proportional to a randomly (uniformly) selected integers ranging from one to 100. Experiments were performed so that there were 100 trials for each value M of the number of bins, which when considering the varying numbers of bins, amounted to $100 \times 100 = 10000$ trials in total. There were no constraints placed on the probability of a bin given the probabilities of the bins adjacent to it. For this reason, in some trials it is possible that there exist adjacent bins with equal bin probabilities.

Experiments were performed with $N = 500, 1000$, and 10000 data points. It is important to note that in the case of $N = 500$ data points, the bins are typically well populated for small numbers of bins, M, since the average number of data points per bin will be proportional to N/M. However, for large numbers of bins, such as $M > 10$, the number of samples per bin can become relatively small, even in relatively probable bins, making accurate inferences difficult. The situation improves for large numbers of bins as the number of samples, N, increases. For this reason, it is to be expected that the performances of the algorithms will change with an increasing number N of data points.

The results for $N = 500, 1000$, and 10000 are listed in Tables 1, 2, and 3, respectively. In all three cases, the proposed OPTBINS method (Knuth) has the lowest overall root mean square error in the bin number estimates. For datasets sampled from distributions with a small number of bins, the OPTBINS method almost always estimated the number of bins correctly. Specifically, in the case of $N = 500$, the OPTBINS method was the only method to correctly identify the number of bins 100% of the time for $M = 1$ to 5. The situation improved for increasing N for small numbers of bins ranging from $M = 1$ to 8, or 9.

Overall, for $N = 500$ data points, AIC correctly identified the number of bins correctly 59% of the time. The Stone method came in second with a 54% success rate. While the OPTBINS method came in third in terms of obtaining the correct number of bins 48% of the time, it came in first in terms of the root-mean-square (RMS) error, which was only 2.34. The other methods had a greater RMS error by an order of magnitude.

For $N = 1000$ data points, the OPTBINS method correctly estimated the number of bins 58% of the time, which again came in third behind AIC and Stone. However, the OPTBINS RMS error was reduced to 1.43, which was at least 8 times better than

Table 2
Summary of the histogram binning performance for $N = 1000$ for data sampled from distributions consisting of the specified Number of Bins. For each algorithm, the fraction of the correct number of bins (cor) is listed as well as the mean (mean) number of bins obtained for the 100 trials. At the bottom, the overall fraction of correct number of bins (cor) is listed along with the root mean square error (rms) in the number of bin estimates for all 10000 trials. The best scores are highlighted in boldface font.

Number of bins	**OPTBINS**		AIC		BIC		Scott		Stone	
	cor	mean	cor	mean	cor	mean	cor	mean	cor	mean
1	**1**	**1**	0.65	2.6	**1**	**1**	0	33.8	0.65	2.5
2	**1**	**2**	0.71	3.9	0.84	2.4	0	23.4	0.67	4.2
3	**1**	**3**	0.73	5.2	0.91	3.2	0	14	0.71	5.3
4	**1**	**4**	0.79	6.4	0.89	**4**	0	10	0.75	6
5	0.92	**4.9**	0.83	6.7	**0.98**	**5.1**	0.13	7.6	0.81	7.1
6	**1**	**6**	0.82	8.7	0.9	5.9	0.22	5.7	0.84	8.6
7	**1**	**7**	0.83	9.8	0.93	7.5	0.06	4.8	0.82	9.8
8	**1**	**8**	0.86	10.1	0.92	**8**	0	4.1	0.83	10.8
9	0.54	8.5	0.85	11.2	**0.92**	**8.9**	0	3.6	0.84	11.6
10	**0.99**	**10**	0.83	13.4	0.92	**10**	0	3.1	0.83	13
20	**0.99**	**20**	0.87	23.3	0.88	19.9	0	1.1	0.86	23
25	0.25	24.3	**0.94**	26.3	0.93	**24.6**	0	1	**0.94**	26.2
30	0.36	**28.7**	**0.86**	33.2	0.81	27.2	0	0.9	0.84	33.1
40	0.35	**38.1**	0.85	43.3	0.63	31	0	0	**0.87**	42.1
50	**0.82**	**49.5**	0.79	55.1	0.33	26.6	0	0	0.79	53.9
60	0.63	**59.3**	**0.81**	62.4	0.19	22.3	0	0	0.8	62.3
70	0.4	68.2	**0.79**	71.2	0.03	13	0	0	0.74	**71.1**
80	0.39	76.3	**0.74**	**80.6**	0	8.6	0	0	0.72	**80.6**
90	0.15	89.2	**0.73**	**90**	0	5.8	0	0	0.7	87.1
100	**0.73**	**98.9**	0.72	97.6	0	4.6	0	0	0.66	94.8
	cor	rms	cor	rms	cor	rms	cor	rms	cor	rms
overall	0.58	**1.43**	**0.80**	8.88	0.43	48.77	0.004	58.19	0.79	8.81

Table 3
Summary of the histogram binning performance for $N = 10000$ for data sampled from distributions consisting of the specified Number of Bins. For each algorithm, the fraction of the correct number of bins (cor) is listed as well as the mean (mean) number of bins obtained for the 100 trials. At the bottom, the overall fraction of correct number of bins (cor) is listed along with the root mean square error (rms) in the number of bin estimates for all 10000 trials. The best scores are highlighted in boldface font.

Number of bins	**OPTBINS**		AIC		BIC		Scott		Stone	
	cor	mean	cor	mean	cor	mean	cor	mean	cor	mean
1	**1**	**1**	0.63	2.6	**1**	**1**	0	73.6	0.62	2.7
2	**1**	**2**	0.65	4.7	0.97	**2**	0	49.2	0.67	3.5
3	**1**	**3**	0.79	5.3	0.99	**3**	0	31.5	0.84	4.5
4	0.78	**3.8**	0.78	9.4	**0.97**	**4.2**	0	22.7	0.79	7.3
5	**1**	**5**	0.85	7.3	**1**	**5**	0	17	0.87	6
6	**1**	**6**	0.89	8.1	0.99	**6**	0	13.5	0.93	6.5
7	0.66	**6.7**	0.91	9.3	**0.99**	7.6	0.03	11.8	0.91	9.1
8	**1**	**8**	0.92	10.6	0.99	**8**	0.16	9.7	0.93	9.7
9	**1**	**9**	0.91	12.6	**1**	**9**	0.2	8.1	0.91	11.5
10	0.34	9.3	0.95	12	**0.99**	**10.2**	0.04	7.3	0.93	11.6
20	0.46	19.5	0.95	23.2	**1**	**20**	0	3.3	0.97	21.6
25	0.93	**24.9**	0.96	27.4	**0.99**	25.3	0	2.5	0.97	26.5
30	0.46	28.4	0.97	32	**1**	**30**	0	2	0.99	30.7
40	0.94	39.9	0.98	40.8	**1**	**40**	0	1.2	0.99	40.4
50	0.78	49.6	0.97	51.5	**1**	**50**	0	1	0.97	51.5
60	0.97	59.9	0.99	60.4	**1**	**60**	0	1	0.99	60.4
70	0.66	69	0.97	70.7	**0.98**	**70.1**	0	0.9	0.97	70.7
80	0.47	78.9	**0.98**	80.3	**0.98**	**80.1**	0	0.2	**0.98**	80.3
90	0.62	89.2	**1**	**90**	**1**	**90**	0	0	**1**	**90**
100	0.73	99.7	**0.99**	**100**	**0.99**	**100**	0	0	**0.99**	**100**
	cor	rms	cor	rms	cor	rms	cor	rms	cor	rms
overall	0.61	**1.26**	0.96	7.90	**0.99**	2.62	0.004	58.32	0.96	6.25

AIC and Stone, and about 50 times better than BIC and Scott. At $N = 1000$, the performance of both AIC and BIC was observed to improve significantly for numbers of bins less than 30.

Last, for $N = 10000$ data points, BIC took the lead as the method that most often obtained the correct number of bins with an overall accuracy of 99%. In contrast, the OPTBINS method correctly estimated the number of bins 61% of the time, which was not as good as both AIC and Stone, which were correct 96% of the time. That being said, the OPTBINS method still exhibited the lowest RMS error, of 1.26, in its bin estimates.

In summary, for small numbers of bins, the OPTBINS method provides correct estimates of bin numbers more reliably than the other methods. In all cases, the OPTBINS method provides estimates that minimize the RMS error. However, in terms of general accuracy, both AIC and the Stone method work best for $N \leq 1000$ data points. For a large number of data points, $N = 10000$, BIC begins to dominate in terms of accuracy for all numbers of bins. The overall benefits of the OPTBINS method is that it exhibits high accuracy for small numbers of bins, which makes it good for testing uniformity, and generally provides estimates that minimize the RMS error.

10. Algorithmic implementations

The basic OPTBINS algorithm takes a one-dimensional data set and performs a brute force exhaustive search that computes the

relative log posterior for all the bin values from 1 to M. An exhaustive search will be slow for large data sets that require large numbers of bins, or multi-dimensional data sets that have multiple bin dimensions. In the case of one-dimensional data, the execution time of the `Matlab` implementation (see Appendix) on a Dell Latitude D610 laptop with an Intel Pentium M 2.13 GHz processor bilinear in both the number of data points N and the number of bins M to consider. The execution time can be estimated by the approximate formula:

$$T = 0.0000171N \cdot M - 0.00026N - 0.0026M, \tag{64}$$

where T is the execution time in seconds. For instance, for $N = 25000$ data values and $M = 500$ bins to consider from $M = 1$ to $M = 500$, the observed time was 194 seconds, which is close to the approximate time of 206 seconds. The algorithm is much faster for small numbers of data points. For $N = 1000$ data points and $M = 50$ bins, the execution time is approximately 0.5 seconds.

There are many techniques that are faster than an exhaustive search. For example, sampling techniques, such as nested sampling [28], are particularly efficient given that there are a finite number of bins to consider. At this point we have implemented the `OPTBINS` model and posterior in the nested sampling framework. The code has been designed to provide the user flexibility in choosing the mode or the mean, which may be more desirable in cases of extremely skewed posterior probabilities. Such a sampling method has the distinct advantage that the entire space is not searched. Moreover, we have added code that stores up to 10000 log posterior results from previously examined numbers of bins and allows us to access them later, resulting in far fewer evaluations, especially in high-dimensional spaces.

The computational bottleneck lies in the binning algorithm that must bin the data in order to compute the factors for the log posterior. We have a `Matlab mex` file implementation, which is essentially compiled `C` code, that speeds up the evaluations an order of magnitude. However, the execution time is still constrained by this step, which depends on both the number data points and the number of bins. The nested sampling algorithm limits the search space by choosing the maximum number of bins in any dimension to be $5N^{1/3}$, which on the average is an order of magnitude greater than the number of bins suggested by Scott's Rule (3) Since in one-dimension the execution time of the brute force algorithm is bilinear in N and M, the execution time should go as

$$Time \propto N \cdot M \sim M^{\frac{D+3}{3}}. \tag{65}$$

We have verified this theoretical estimate with tests that show times increasing as $M^{1.4}$ for one-dimensional data, $M^{1.6}$ for two-dimensional data, and $M^{2.0}$ for three-dimensional data, which compare reasonably well with the predicted exponents of 4/3, 5/3, and 7/3 for one, two and three dimensions respectively. The advantage of the nested sampling algorithm over an exhaustive search arises from a significant reduction in the number of calls to the binning algorithm. This is both due to the fact that nested sampling does not search the entire space, and the fact that the log probabilities of previous computations are stored for easy lookup in the event that a model with the particular number of bins is visited multiple times. In one-dimension, nested sampling visits the majority of the bins and does not significantly outperform an exhaustive search. However in two and three dimensions, nested sampling visits successively fewer bin configurations, which results in remarkable speed-ups over the exhaustive search algorithm. For example, in three-dimensions, exhaustive search takes 2822 seconds for 2000 data points; whereas nested sampling takes only 480 seconds.

The exhaustive search `OPTBINS` algorithm and the nested sampling implementation and supporting code in `Matlab` can be downloaded from: https://github.com/khknuth/histo. The `OPTBINS` algorithm has been implemented in Mathematica's *DensityHistogram* function as the *Knuth Binning Method*. It has also been coded into Python for *AstroML* (http://astroml.github.com/) under the function name `knuth_nbins` where it is referred to as *Knuth's Rule* [29,30]. *AstroML* is a freely available Python repository for tools and algorithms commonly used for statistical data analysis and machine learning in astronomy and astrophysics.

11. Discussion

The optimal binning algorithm, `OPTBINS`, also known as the Knuth method, presented in this paper relies on finding the mode of the marginal posterior probability of the number of bins in a piecewise-constant density function model of the distribution from which the data were sampled. This posterior probability originates as a product of the likelihood of the density parameters given the data and the prior probability of those same parameter values, where the prior probability is a non-informative prior. As the number of bins increases, the prior probability (14), which depends on the inverse of the square root of the product of the bin probabilities, tends to increase. Meanwhile, the joint likelihood (10), which is a product of the bin probabilities of the individual data tends to decrease.[2] Since the posterior is a product of these two functions, the maximum of the posterior probability occurs at a point where these two opposing factors are balanced. This interplay between the likelihood and the prior probability effectively implements Occam's razor by selecting the most simple model that best describes the data.

We have studied this algorithm's ability to model the underlying density by comparing its behavior to several other popular bin selection techniques: Akaike model selection criterion (AIC) [24], the Bayesian Information Criterion (BIC) [25,26], Stone's Rule [4], and Scott's rule [1,2], which is similar to the rule proposed by Freedman and Diaconis [3]. In terms of general accuracy, both AIC and the Stone method work best for $N \leq 1000$ data points. For a large number of data points, $N = 10000$, BIC began to dominate in terms of accuracy for all numbers of bins. The `OPTBINS` algorithm exhibited the highest accuracy for small numbers of bins, which makes it good for testing uniformity, and generally provided estimates that minimize the RMS error.

The utility of this algorithm was also demonstrated by applying it to three real data sets. In two of the three cases the algorithm recommended reasonable bin numbers. In the third case involving the Old Faithful data set it revealed that the data were excessively rounded. That is, the discrete nature of the data was a more salient feature than the shape of the underlying density function. To obtain a reasonable number of bins in this case, one need only add sufficiently small random numbers to the original data points. However, in the example of the Old Faithful data set `OPTBINS` indicates something more serious. Excessive rounding has resulted in data of poor quality and this may have had an impact on previous studies. The fact that the `OPTBINS` algorithm can identify data sets where the data have been excessively rounded may be of benefit in identifying problem data sets as well as selecting an appropriate degree of rounding in cases where economic storage or transmission are an issue [31].

In addition to these applications, `OPTBINS` already has been used to generate histograms in several other published studies. One study by Nir et al. [32] involved making histograms of rational numbers, which led to particularly pathological 'spike and

[2] This is the reverse from what one usually expects where increasing the number of parameters decreases the prior and increases the likelihood.

void' distributions. In this case, there is no maximum to the log posterior. However, adding small random numbers to the data, as recommended by Bayman and Broadhurst [17], was again shown to be an effective remedy.

Our algorithm also can be readily applied to multi-dimensional data sets, which we demonstrated with a two-dimensional data set. In practice, we have been applying OPTBINS to three-dimensional data sets with comparable results. We have also implemented a nested sampling algorithm that enables the user to select either the most probable number of bins (mode) or the mean number of bins. The nested sampling implementation displays significant speed-up over the exhaustive search algorithm, especially in the case of higher dimensions.

It should be noted that we are working with a piecewise-constant model of the density function, and *not* a histogram *per se*. The distinction is subtle, but important. Given the full posterior probability for the model parameters and a selected number of bins, one can estimate the mean bin probabilities and their associated standard deviations. This is extremely useful in that it quantifies uncertainties in the density model, which can be used in subsequent calculations. In this paper, we demonstrated that with small numbers of data points the magnitude of the error bars on the bin heights is on the order of the bin heights themselves. Such a situation indicates that too few data exist to infer a density function. This can also be determined by examining the marginal posterior probability for the number of bins. In cases where there are too few data points, the posterior will not possess a well-defined mode. In our experiments with Gaussian-distributed data, we found that approximately 150 data points are needed to accurately estimate the density model when the functional form of the density is unknown.

We have made some simplifying assumptions in this work. First, the data points themselves are assumed to have no associated uncertainties. Second, the endpoints of the density model are defined by the extreme data values, and are not allowed to vary during the analysis. Third, we use the marginal posterior to select the optimal number of bins and then use this value to estimate the mean bin heights and their variance. This neglects uncertainty about the number of bins, which means that the variance in the bin heights is underestimated.

Equal width bins can be very inefficient in describing multimodal density functions (as in Fig. 1G.) In such cases, variable bin-width models such as the maximum likelihood estimation introduced by Wegman [6], Bayesian partitioning [7], Bayesian Blocks [8], Bayesian bin distribution inference [9], Bayesian regression of piecewise constant functions [10], and Bayesian model determination through techniques such as reversible jump Markov chain Monte Carlo [33] may be more appropriate options in certain research applications.

For many applications, OPTBINS efficiently delivers histograms with a number of bins that provides an appropriate depiction of the shape of the density function given the available data while minimizing the appearance of random fluctuations. While this algorithm produces the most probable piecewise-constant model of the density function, it is not assured that parameter estimates based on this density model will be optimal. This simply due to the way in which probability densities transform under changes of variables. Given the most probable value $\hat{x}$ of parameter x, it is not always true that $\hat{y} = f(\hat{x})$ will be the most probable value of the parameter $y = f(x)$. For this reason, parameters derived from this optimal density model are not guaranteed to be optimal.

A Matlab implementation of the basic algorithm is given in the Appendix and the OPTBINS package can be downloaded from https://github.com/khknuth/histo. A Python implementation is available from *AstroML* (http://astroml.github.com/) under the function name knuth_nbins where it is referred to as *Knuth's Rule* [29]. Last, this algorithm has been implemented in Mathematica's *DensityHistogram* function as the *Knuth Binning Method*.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author thanks the NASA Earth-Sun Systems Technology Office Applied Information Systems Technology Program and the NASA Applied Information Systems Research Program for their support through the grants NASA AISRP NNH05ZDA001N and NASA ESTC NNX07AD97A and NNX07AN04G. The author is grateful for many valuable conversations, interactions, comments and feedback from Amy Braverman, John Broadhurst, J. Pat Castle, Joseph Coughlan, Charles Curry, Deniz Gencaga, Karen Huyser, Ercan Kuruoglu, Raquel Prado, Carlos Rodriguez, William Rossow, Jeffrey Scargle, Devinder Sivia, John Skilling, Len Trejo, Jacob Vanderplas, Michael Way, Kevin Wheeler, David Wolpert, and especially the enthusiastic support of Barry Stokes. The author is also grateful to an anonymous reviewer who caught an error in a previous version of this manuscript. Thanks also to Anthony Gotera who coded the histogram binning mex file and Yangxun (Billy) Chen who helped to code the excessive rounding detection subroutine. Thanks also to the creators of AstroML (Zeljko Ivezic, Andrew Connolly, Jacob Vanderplas, and Alex Gray) who have made this algorithm freely available in the AstroML Python repository. Two data sets used in this paper were obtained with the assistance of UCI Repository of machine learning databases [22] and the kind permission of Warwick Nash and Therese Stukel. The Old Faithful data set was made available by Education Queensland and maintained by Rex Boggs at: http://exploringdata.net/datasets.htm

Appendix. Matlab code

```
% OPTBINS finds the optimal number of bins for a one-dimensional
% data set using the posterior probability for the number of bins
% This algorithm uses a brute-force search trying every possible
% bin number in the given range.  This can of course be improved.
% Generalization to multidimensional data sets is straightforward.
%
% Usage:
%           optM = OPTBINS(data,maxM);
% Where:
%           data is a (1,N) vector of data points
%           maxM is the maximum number of bins to consider
%
% Cite: K.H. Knuth. Optimal data-based binning for histograms
% and histogram-based probability density models, Digital Signal
% Process. (2019) https://10.1016/j.dsp.2019.102581

function optM = OPTBINS(data,maxM)

if size(data)>2 | size(data,1)>1
    error('data dimensions must be (1,N)');
end
N = size(data,2);

% Simply loop through the different numbers of bins
% and compute the posterior probability for each.
logp = zeros(1,maxM);
for M = 1:maxM
    n = hist(data,M);  % Bin the data (equal width bins here)
    part1 = N*log(M) + gammaln(M/2) - gammaln(N+M/2);
    part2 = - M*gammaln(1/2) + sum(gammaln(n+0.5));
    logp(M) = part1 + part2;
end

[maximum, optM] = max(logp);
return;
```

References

[1] D.W. Scott, On optimal and data-based histograms, Biometrika 66 (1979) 605–610.
[2] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 1992.
[3] D. Freedman, P. Diaconis, On the histogram as a density estimator: l_2 theory, Z. Wahrscheinlichkeitstheor. Verw. Geb. 57 (1981) 453–476.
[4] C.J. Stone, An asymptotically histogram selection rule, in: J. Neyman (Ed.), Proc. Second Berkeley Symp., Univ. California Press, Berkeley, 1984, pp. 513–520.
[5] M. Rudemo, Empirical choice of histograms and kernel density estimators, Scand. J. Stat. 9 (1982) 65–78.
[6] E.J. Wegman, Maximum likelihood estimation of a probability density function, Sankhyā: Ind. J. Stat. 37 (1975) 211–224.
[7] D.G.T. Denison, N.M. Adams, C.C. Holmes, D.J. Hand, Bayesian partition modelling, Comput. Stat. Data Anal. 38 (4) (2002) 475–485.
[8] B. Jackson, J. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, T.T. Tsai, An algorithm for optimal partitioning of data on an interval, IEEE Signal Process. Lett. 12 (2005) 105–108.
[9] A. Endres, P. Foldiak, Bayesian bin distribution inference and mutual information, IEEE Trans. Inf. Theory 51 (11) (2005) 3766–3779.
[10] M. Hutter, Exact Bayesian regression of piecewise constant functions, Bayesian Anal. 2 (4) (2007) 635–664.
[11] H. Jeffreys, Theory of Probability, 3rd edition, Oxford University Press, Oxford, 1961.
[12] G.E.P. Box, G.C. Tiao, Bayesian Inference in Statistical Analysis, John Wiley & Sons, New York, 1992.
[13] J.O. Berger, J.M. Bernardo, Ordered group reference priors with application to the multinomial problem, Biometrika 79 (1992) 25–37.
[14] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, Berlin, 2006.
[15] M. Hutter, Distribution of mutual information, in: T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14, The MIT Press, Cambridge, 2001, pp. 399–406.
[16] M. Abramowitz, I.A. Stegun, Handbook of Mathematical Functions, Dover Publications, Inc., New York, 1972.
[17] B.F. Bayman, J.B. Broadhurst, A simple solution to a problem arising from the processing of finite accuracy digital data using integer arithmetic, Nucl. Instrum. Methods 167 (1979) 475–478.
[18] G. Arfken, Mathematical Methods for Physicists, Academic Press, Orlando, FL, 1985.
[19] W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn, W.B. Ford, The Population Biology of Abalone (Haliotis Species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait, Tech. Rep. 48, Sea Fisheries Division, Marine Research Laboratories-Taroona, Department of Primary Industry and Fisheries, Tasmania, 1994.
[20] D.W. Nierenberg, T.A. Stukel, J.A. Baron, B.J. Dain, E.R. Greenberg, Determinants of plasma levels of beta-carotene and retinol, Am. J. Epidemiol. 130 (1989) 511–521.
[21] A. Azzalini, A.W. Bowman, A look at some data on the Old Faithful Geyser, Appl. Stat. 39 (1990) 357–365.
[22] D.J. Newman, S. Hettich, C.L. Blake, C.J. Merz, UCI repository of machine learning databases, http://www.ics.uci.edu/~mlearn/MLRepository.html, 1998.
[23] R.W. West, W. Ogden, Interactive demonstrations for statistics education on the world wide web, J. Stat. Educ. 6 (3) (1998), http://www.stat.sc.edu/~west/javahtml/Histogram.html.
[24] H. Akaike, A new look at the statistical identification model, IEEE Trans. Autom. Control 19 (1974) 716–723.
[25] G. Schwarz, Estimating the dimension of a model, Ann. Stat. 6 (2) (1978) 461–464.
[26] E. Wit, E.v.d. Heuvel, J.-W. Romeijn, 'all models are wrong...': an introduction to model uncertainty, Stat. Neerl. 66 (3) (2012) 217–236.
[27] J.A. Hartigan, Bayesian histograms, in: J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Eds.), Bayesian Statistics, vol. 5, Oxford Univ. Press, Oxford, 1996, pp. 211–222.
[28] D.S. Sivia, J. Skilling, Data Analysis. A Bayesian Tutorial, 2nd edition, Oxford University Press, Oxford, 2006.
[29] Z. Ivezić, A.J. Connolly, J. Vanderplas, A. Gray, Statistics, Data Mining and Machine Learning in Astronomy, Princeton University Press, 2013.
[30] J.T. Vanderplas, A.J. Connolly, Ž. Ivezić, A. Gray, Introduction to astroML: machine learning for astrophysics, in: Conference on Intelligent Data Understanding (CIDU), NASA, 2012, pp. 47–54.
[31] K.H. Knuth, J.P. Castle, K.R. Wheeler, Identifying excessively rounded or truncated data, in: A. Rizzi, M. Vichi (Eds.), Proc. of the 17th Meeting of the Int. Assoc. for Statistical Computing-European Regional Section: Computational Statistics (COMPSTAT 2006), Physica-Verlag, Heidelberg, 2006, pp. 313–324.
[32] E. Nir, X. Michalet, K. Hamadani, T.A. Laurence, D. Neuhauser, Y. Kovchegov, S. Weiss, Shot-noise limited single-molecule FRET histogram: comparison between theory and experiments, J. Phys. Chem. B 110 (2006) 22103–22124.
[33] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732.

Kevin H. Knuth is an Associate Professor in the Department of Physics at the University at Albany, Albany NY USA. He is Editor-in-Chief of the journal Entropy, and is a former NASA research scientist. He has over 20 years of experience in applying Bayesian and maximum entropy methods to the design of machine learning algorithms for data analysis applied to astronomy and the physical sciences. His current research interests include the foundations of physics, autonomous robotics, and searching for and characterizing extrasolar planets.