

Version 22: Draft of 2006-02-21: 17:54

Archived at <http://www.huginn.com/knuth/papers/knuth-histo-draft-060221.pdf>

## Optimal Data-Based Binning for Histograms

Kevin H. Knuth

Department of Physics, University at Albany (SUNY), Albany, NY, 12222, USA

*knuth@albany.edu*

Received \_\_\_\_\_; accepted \_\_\_\_\_

## 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. In this paper we introduce a straightforward data-based method of determining the optimal number of bins in a uniform bin-width histogram. Using the Bayesian framework, we derive the posterior probability for the number of bins in the density model given the data. The most probable solution is determined naturally by a balance between the likelihood function, which increases with increasing number of bins, and the prior probability of the model, which decreases with increasing number of bins. We demonstrate how these results outperform several well-accepted rules for choosing bin sizes even in the integrated square error sense. Last, we demonstrate that these results can be applied directly to multi-dimensional histograms.

## 1. Introduction

Histograms are used extensively as nonparametric 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 5-20 bins is usually adequate (for example, `Matlab` uses 10 bins as a default). Scott(1979) and Freedman and Diaconis(1981) 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  $f(x)$ ,

$$L(h(x), f(x)) = \int (h(x) - f(x))^2. \quad (1)$$

For  $N$  data points, the optimal bin width  $v$  goes as  $\alpha N^{-1/3}$ , where  $\alpha$  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  $\alpha = 3.49s$  (Scott, 1979), and

$$v_{\text{scott}} = 3.49sN^{-1/3}. \quad (2)$$

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

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

Freedman and Diaconis report similar results, however they suggest choosing  $\alpha$  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, which are often seen in dynamical systems. This is because they were derived assuming particular characteristics of the underlying density. In particular, the result 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  $\int f'^2 > 0$ .

Stone(1984) chooses to minimize  $L(h, f) - \int f^2$  and obtains a rule where one chooses the bin width  $v$  to minimize

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

where  $M$  is the number of bins and  $\pi_i$  are the bin probabilities. Rudemo(1982) obtains a similar rule by applying cross-validation techniques with a Kullback-Leibler risk function.

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 the true density. Instead, we consider 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. This enables one to objectively select an optimal piecewise-constant model describing the density function from which the data were sampled.

## 2. The Piecewise-Constant Density Model

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 width  $v_k$ , where  $k$  is used to index the bins. We further assume that the bins have equal width  $v = v_k$  for all  $k$ , and together they encompass an entire width  $V = Mv$ .<sup>1</sup> 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 total probability mass of  $\pi_k = h_k v_k$  for the bin. This leads to the following piecewise-constant model  $h(x)$  of the unknown probability density function  $f(x)$

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

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

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

Our density model can be re-written in terms of the bin probabilities  $\pi_k$  as

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

Given  $M$  bins and the normalization condition that the integral of the probability density equals unity, we are left with  $M - 1$  bin probabilities:  $\pi_1, \pi_2, \dots, \pi_{M-1}$ , 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

---

<sup>1</sup>For a one-dimensional histogram,  $v_k$  is the width of the  $k^{\text{th}}$  bin. In the case of a multi-dimensional histogram, this will be a multi-dimensional volume.

simplicity, we assume that the bin alignment is fixed so that extreme data points lie precisely at the center of the extreme bins of the histogram (that is, the smallest sample is at the center of the leftmost bin, and similarly for the largest sample). As we will show, this technique is easily extended to multi-dimensional densities of arbitrarily high dimension.

### 2.1. The Likelihood of the Piecewise-Constant Model

The likelihood that a data point  $d_n$  found to be in the  $k^{th}$  bin could have been drawn from a density function described by our model is simply

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

where  $I$  represents our prior knowledge about the problem, such as the range of the data and the bin alignment. For equal width bins, this reduces to

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

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

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

where  $\underline{d} = \{d_1, d_2, \dots, d_N\}$  and  $\underline{\pi} = \{\pi_1, \pi_2, \dots, \pi_{M-1}\}$ . 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 point  $d$ , the likelihood that it will take the value  $x$  is

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

Integrating over all possible values of  $x$ , and writing the bin width as  $v = \frac{V}{M}$

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

$$= \frac{1}{v} \sum_{k=1}^M \int_{-\infty}^{\infty} dx \pi_k \Pi(x_{k-1}, x, x_k) \quad (13)$$

$$= \frac{1}{v} \sum_{k=1}^M \pi_k v \quad (14)$$

$$= \sum_{k=1}^M \pi_k \quad (15)$$

$$= 1. \quad (16)$$

## 2.2. The Prior Probabilities

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

$$p(M|I) = \begin{cases} C^{-1} & \text{if } 1 \leq M \leq C \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

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  $\pi_1, \pi_2, \dots, \pi_{M-1}$ , the possible values of which lie within a simplex defined by the corners of an  $M$ -dimensional hypercube with unit side lengths

$$p(\underline{\pi}|M, I) = \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \left[ \pi_1 \pi_2 \cdots \pi_{M-1} \left( 1 - \sum_{i=1}^{M-1} \pi_i \right) \right]^{-1/2}. \quad (18)$$

Equation (18) is the Jeffreys's prior for the multinomial likelihood (10) Jeffreys(1961), Box and Tiao(1992), Berger and Bernardo(1992), and has the advantage in that it is also the conjugate prior to the multinomial likelihood.

## 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

$$p(\underline{\pi}, M|\underline{d}, I) \propto p(\underline{\pi}|I) p(M|I) p(\underline{d}|\underline{\pi}, M, I). \quad (19)$$

Substituting (10), (17), and (18) gives the joint posterior probability for the piecewise-constant density model

$$p(\underline{\pi}, M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \pi_1^{n_1-\frac{1}{2}} \pi_2^{n_2-\frac{1}{2}} \dots \pi_{M-1}^{n_{M-1}-\frac{1}{2}} \left(1 - \sum_{i=1}^{M-1} \pi_i\right)^{n_M-\frac{1}{2}}, \quad (20)$$

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 complete posterior over all possible values of  $\pi_1, \pi_2, \dots, \pi_{M-1}$  in the simplex. The expression we desire is written as a series of nested integrals over the  $M-1$  dimensional parameter space of bin probabilities

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \int_0^1 d\pi_1 \pi_1^{n_1-\frac{1}{2}} \int_0^{1-\pi_1} d\pi_2 \pi_2^{n_2-\frac{1}{2}} \dots \dots \int_0^{(1-\sum_{i=1}^{M-2} \pi_i)} d\pi_{M-1} \pi_{M-1}^{n_{M-1}-\frac{1}{2}} \left(1 - \sum_{i=1}^{M-1} \pi_i\right)^{n_M-\frac{1}{2}}. \quad (21)$$

In order to write this more compactly, we first define

$$\begin{aligned} a_1 &= 1 \\ a_2 &= 1 - \pi_1 \\ a_3 &= 1 - \pi_1 - \pi_2 \\ &\vdots \\ a_{M-1} &= 1 - \sum_{k=1}^{M-2} \pi_k \end{aligned} \quad (22)$$

and note the recursion relation

$$a_k = a_{k-1} - \pi_{k-1}. \quad (23)$$

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

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \int_0^{a_1} d\pi_1 \pi_1^{n_1-\frac{1}{2}} \int_0^{a_2} d\pi_2 \pi_2^{n_2-\frac{1}{2}} \dots \dots \int_0^{a_{M-1}} d\pi_{M-1} \pi_{M-1}^{n_{M-1}-\frac{1}{2}} (a_{M-1} - \pi_{M-1})^{n_M-\frac{1}{2}}. \quad (24)$$

To solve the set of nested integrals in (21), consider the general integral

$$I_k = \int_0^{a_k} d\pi_k \pi_k^{n_k-\frac{1}{2}} (a_k - \pi_k)^{b_k} \quad (25)$$

where  $b_k \in \mathbb{R}^+$  and  $b_k > 1/2$ . This integral can be re-written as

$$I_k = a_k^{b_k} \int_0^{a_k} d\pi_k \pi_k^{n_k-\frac{1}{2}} \left(1 - \frac{\pi_k}{a_k}\right)^{b_k}. \quad (26)$$

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

$$\begin{aligned}
 I_k &= a_k^{b_k} \int_0^1 du a_k^{n_k + \frac{1}{2}} u^{n_k - \frac{1}{2}} (1-u)^{b_k} \\
 &= a_k^{b_k + n_k + \frac{1}{2}} \int_0^1 du u^{n_k - \frac{1}{2}} (1-u)^{b_k}, \\
 &= a_k^{b_k + n_k + \frac{1}{2}} B\left(n_k + \frac{1}{2}, b_k + 1\right)
 \end{aligned} \tag{27}$$

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

$$B\left(n_k + \frac{1}{2}, b_k + 1\right) = \frac{\Gamma\left(n_k + \frac{1}{2}\right)\Gamma(b_k + 1)}{\Gamma\left(n_k + \frac{1}{2} + b_k + 1\right)}. \tag{28}$$

To solve all of the integrals we rewrite  $a_k$  in (27) using the recursion formula (23)

$$I_k = (a_{k-1} - \pi_{k-1})^{b_k + n_k + \frac{1}{2}} B\left(n_k + \frac{1}{2}, b_k + 1\right). \tag{29}$$

By defining

$$\begin{aligned}
 b_{M-1} &= n_M - \frac{1}{2} \\
 b_{k-1} &= b_k + n_k + \frac{1}{2}
 \end{aligned} \tag{30}$$

we find

$$b_1 = N - n_1 + \frac{M}{2} - \frac{3}{2}. \tag{31}$$

Finally, integrating (24) gives

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \prod_{k=1}^{M-1} B\left(n_k + \frac{1}{2}, b_k + 1\right), \tag{32}$$

which can be simplified further by expanding the Beta functions using (28)

$$\begin{aligned}
 p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} &\frac{\Gamma\left(n_1 + \frac{1}{2}\right)\Gamma(b_1 + 1)}{\Gamma\left(n_1 + \frac{1}{2} + b_1 + 1\right)} \cdot \frac{\Gamma\left(n_2 + \frac{1}{2}\right)\Gamma(b_2 + 1)}{\Gamma\left(n_2 + \frac{1}{2} + b_2 + 1\right)} \cdots \\
 &\cdots \frac{\Gamma\left(n_{M-1} + \frac{1}{2}\right)\Gamma(b_{M-1} + 1)}{\Gamma\left(n_{M-1} + \frac{1}{2} + b_{M-1} + 1\right)}
 \end{aligned} \tag{33}$$

Using the recursion relation (30) 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

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M} \frac{\prod_{k=1}^M \Gamma\left(n_k + \frac{1}{2}\right)}{\Gamma\left(n_1 + b_1 + \frac{3}{2}\right)}, \tag{34}$$

where we have used (30) to observe that  $\Gamma(b_{M-1} + 1) = \Gamma(n_M + 1/2)$ . Last, again using the recursion relation in (30) we find that  $b_1 = N - n_1 + \frac{M}{2} - \frac{3}{2}$ , which results in our simplified posterior probability

$$p(M|\underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}. \quad (35)$$

Rather than working with the posterior probability directly, it is easier to maximize the logarithm of the posterior

$$\begin{aligned} \log p(M|\underline{d}, I) = N \log M + \log \Gamma\left(\frac{M}{2}\right) - M \log \Gamma\left(\frac{1}{2}\right) - \log \Gamma\left(N + \frac{M}{2}\right) + \\ + \sum_{k=1}^M \log \Gamma\left(n_k + \frac{1}{2}\right) + K, \end{aligned} \quad (36)$$

where  $K$  represents the sum of the volume term and the logarithm of the implicit proportionality constant. The optimal number of bins  $\hat{M}$  is found by

$$\hat{M} = \arg \max_M \{\log p(M|\underline{d}, I)\}. \quad (37)$$

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(1972) can be used for integer  $m$ .

$$\log \Gamma(m) = \sum_{k=1}^{m-1} \log k \quad (38)$$

$$\log \Gamma\left(m + \frac{1}{2}\right) = \frac{1}{2} \log \pi - n \log 2 + \sum_{k=1}^m \log(2k - 1) \quad (39)$$

Equation (35) allows one to easily identify the number of bins  $M$  which optimize the posterior. A `Matlab` code implementation is provided in Appendix 1.

### 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 (64) 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 (21) 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

$$p(\underline{\pi}|\underline{d}, M, I) = \frac{p(\underline{\pi}, M|\underline{d}, I)}{p(M|\underline{d}, I)} \quad (40)$$

where the numerator is given by (64) and the denominator by (35). Since the bins can be treated in any order, we derive the marginal posterior for the first bin and generalize the result for the  $k^{th}$  bin. The marginal posterior is

$$p(\pi_1|\underline{d}, M, I) = \frac{\left(\frac{M}{V}\right)^N \frac{\Gamma\left(\frac{M}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^M}}{p(M|\underline{d}, I)} \pi_1^{n_1 - \frac{1}{2}} \int_0^{a_2} d\pi_2 \pi_2^{n_2 - \frac{1}{2}} \int_0^{a_3} d\pi_3 \pi_3^{n_3 - \frac{1}{2}} \dots \int_0^{a_{M-1}} d\pi_{M-1} \pi_{M-1}^{n_{M-1} - \frac{1}{2}} (a_{M-1} - \pi_{M-1})^{n_M - \frac{1}{2}}. \quad (41)$$

Evaluating the integrals and substituting (35) into the denominator we get

$$p(\pi_1|\underline{d}, M, I) = \frac{\prod_{k=2}^M B(n_k + \frac{1}{2}, b_k + 1)}{\prod_{k=1}^M B(n_k + \frac{1}{2}, b_k + 1)} \pi_1^{n_1 - \frac{1}{2}} (1 - \pi_1)^{b_1} \quad (42)$$

Cancelling terms and explicitly writing  $b_1$ , the marginal posterior for  $\pi_1$  is

$$p(\pi_1|\underline{d}, M, I) = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_1 + \frac{1}{2})\Gamma(N - n_1 + \frac{M-1}{2})} \pi_1^{n_1 - \frac{1}{2}} (1 - \pi_1)^{N - n_1 + \frac{M-3}{2}}, \quad (43)$$

which can easily be verified to be normalized by integrating  $\pi_1$  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  $k^{th}$  bin

$$p(\pi_k|\underline{d}, M, I) = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \pi_k^{n_k - \frac{1}{2}} (1 - \pi_k)^{N - n_k + \frac{M-3}{2}}. \quad (44)$$

The mean bin probability mass can be found from its expectation

$$\langle \pi_k \rangle = \int_0^1 d\pi_k \pi_k p(\pi_k|\underline{d}, M, I), \quad (45)$$

which substituting (44) gives

$$\langle \pi_k \rangle = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \int_0^1 d\pi_k \pi_k^{n_k + \frac{1}{2}} (1 - \pi_k)^{N - n_k + \frac{M-3}{2}}. \quad (46)$$

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

$$\langle \pi_k \rangle = \frac{\Gamma(N + \frac{M}{2})}{\Gamma(n_k + \frac{1}{2})\Gamma(N - n_k + \frac{M-1}{2})} \cdot \frac{\Gamma(n_k + \frac{3}{2})\Gamma(N - n_k + \frac{M-1}{2})}{\Gamma(N + \frac{M}{2} + 1)}. \quad (47)$$

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

$$\langle \pi_k \rangle = \frac{n_k + \frac{1}{2}}{N + \frac{M}{2}}. \quad (48)$$

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

$$\mu_k = \langle h_k \rangle = \frac{\langle \pi_k \rangle}{v_k} = \left(\frac{M}{V}\right) \left(\frac{n_k + \frac{1}{2}}{N + \frac{M}{2}}\right). \quad (49)$$

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 variance of the height of the  $k^{\text{th}}$  bin is found similarly by

$$\sigma_k^2 = \left(\frac{M}{V}\right)^2 (\langle \pi_k^2 \rangle - \langle \pi_k \rangle^2), \quad (50)$$

which gives

$$\sigma_k^2 = \left(\frac{M}{V}\right)^2 \left( \frac{(n_k + \frac{1}{2})(N - n_k + \frac{M-1}{2})}{(N + \frac{M}{2} + 1)(N + \frac{M}{2})^2} \right). \quad (51)$$

Thus, given the optimal number of bins found by maximizing (36), the mean and variance of the bin heights are found from (49) and (51), which allow us to construct an explicit histogram model of the probability density and perform computations and error analysis. Note that in the case where there is one bin (51) 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 histogram model of several different data sets. Here we consider 1000 data points sampled from four different probability density functions. The optimal histogram for 1000 data points drawn from a Normal distribution is shown in Figure 1A, where it is superimposed over a 100-bin histogram showing the density of the sampled points. Figure 1B shows that the logarithm of the posterior probability (36) peaks at 9 bins. Figure 1C shows the optimal binning for data sampled from a 4-step piecewise-constant density. The logarithm of the posterior (Figure 1D) peaks at 4 bins, which indicates that the algorithm can correctly detect the 4-step structure. In figures 1E and F, we see 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. Last, we consider a density function that consists of a mixture of three sharply-peaked Gaussians with a uniform background (Figure 1G). The posterior peaks at 52 bins indicating that the data warrant a detailed model (Figure 1H). The spikes in the log posterior are due to the fact that the bin edges are fixed. The 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.

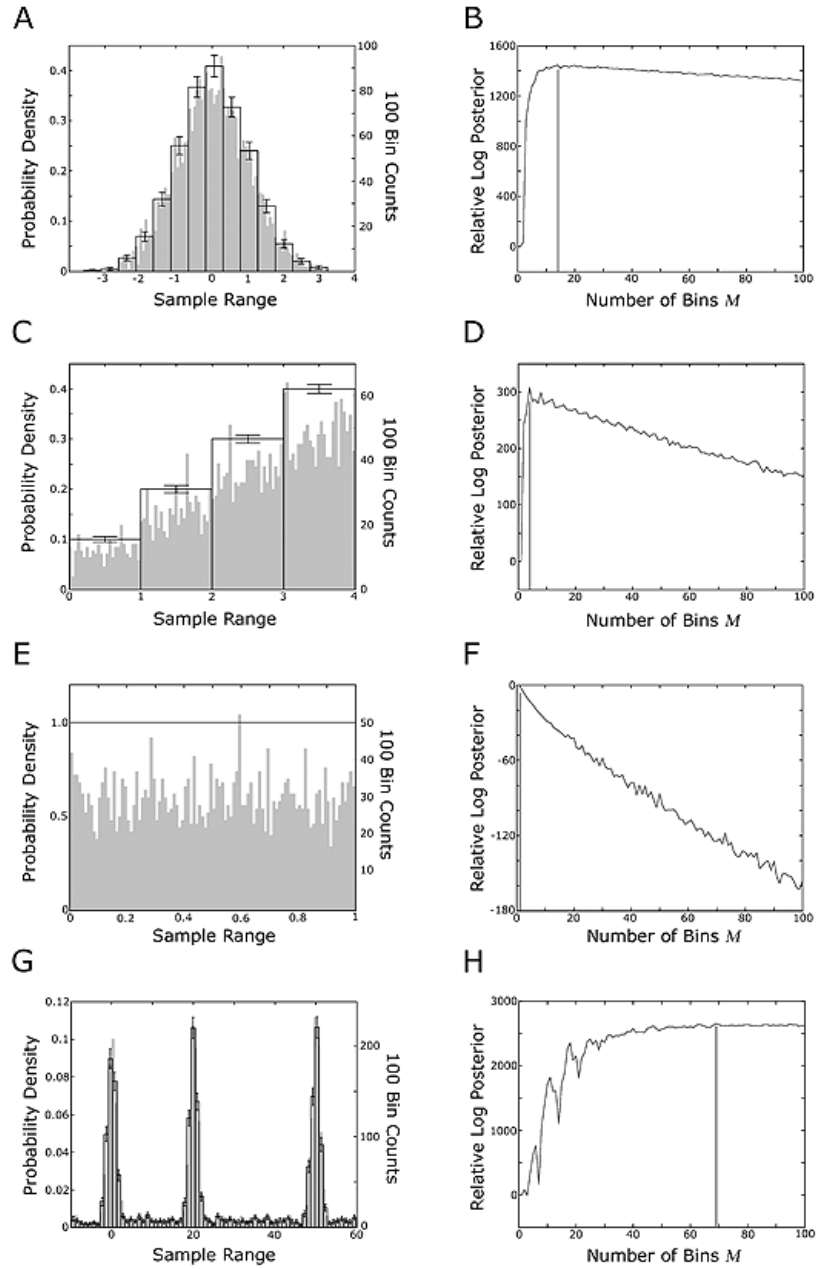


Fig. 1.— To demonstrate the technique, 1000 samples were drawn from four different probability density functions (pdf) above. (A) The optimal histogram for 1000 samples drawn from a Gaussian density function is superimposed over a 100-bin histogram that shows the distribution of data samples. (B) The log posterior probability of the number of bins peaks at nine bins for these 1000 data sampled from the Gaussian density. (C) Samples shown are from a 4-step piecewise constant density function. The optimal binning describes this density accurately since (D) the log posterior peaks at four bins. (E) These data were sampled from a uniform density as verified by the log posterior probability (F). (G) shows 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.

#### 4.2. Comparison to Other Techniques

We now compare our results with those obtained using Scott’s Rule, Stone’s Rule, and the Akaike model selection criterion (Akaike(1974)).<sup>2</sup> Akaike’s method, applied to histograms in Hartigan(1996), balances the logarithm of the likelihood of the model against the number of model parameters. The number of bins is chosen to maximize

$$AIC(M) = \log p(d_n | \pi_k, M, I) - M. \quad (52)$$

Since Scott’s Rule was derived to be asymptotically optimal for Gaussian distributions, we limited our comparison to Gaussian-distributed data. We tested data sets with 14 different numbers of samples including  $N = \{50, 100, 200, 500, 1000, 2000, \dots, 1000000\}$ . For each  $N$  we tested 50 different histograms, each with  $N$  samples drawn from a Gaussian distribution  $\mathcal{N}(0, 1)$ , for a total of 700 histograms in this analysis. The optimal number of bins was found using Scott’s Rule (3), Stone’s Rule (4), Akaike’s AIC (52) and our present `optBINS` algorithm (36). The quality of fit was quantified in two ways. First, we computed the Integrated Square Error (ISE), which is the criterion for which Scott’s Rule is asymptotically optimized. This is

$$ISE = \int_{-\infty}^{\infty} dx \left( h(x, M) - \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \right)^2, \quad (53)$$

where  $h(x, M)$  is the piecewise-constant histogram model with  $M$  bins. Second, we computed the logarithm of the posterior probability (36), and used it to calculate the log-odds ratio, which is the ratio of the log of the posterior probabilities of  $M_{\text{optBINS}}$  and the bin numbers selected by the other three techniques. For example

$$\text{LogOdds} = \frac{\log p(M_{\text{optBINS}} | \underline{d}, I)}{\log p(M_{\text{Scott}} | \underline{d}, I)}. \quad (54)$$

Figure 2A shows the mean number of bins found using Scott’s Rule and `optBINS` over the entire array of values for the number of samples  $N$ . Fifty histograms were analyzed at each value of  $N$ , giving both the mean number of bins and the standard deviation. Relative to `optBINS`, Scott’s Rule consistently overestimates the number of bins necessary for optimal representation of the density function.

---

<sup>2</sup>Since Freedman and Diaconis’ (F&D) method has the same functional form as Scott’s Rule, the results using F&D are not presented here. Their technique leads to a greater number of bins than Scott’s Rule, which, as we will show, already prescribes more bins than are warranted by the data.

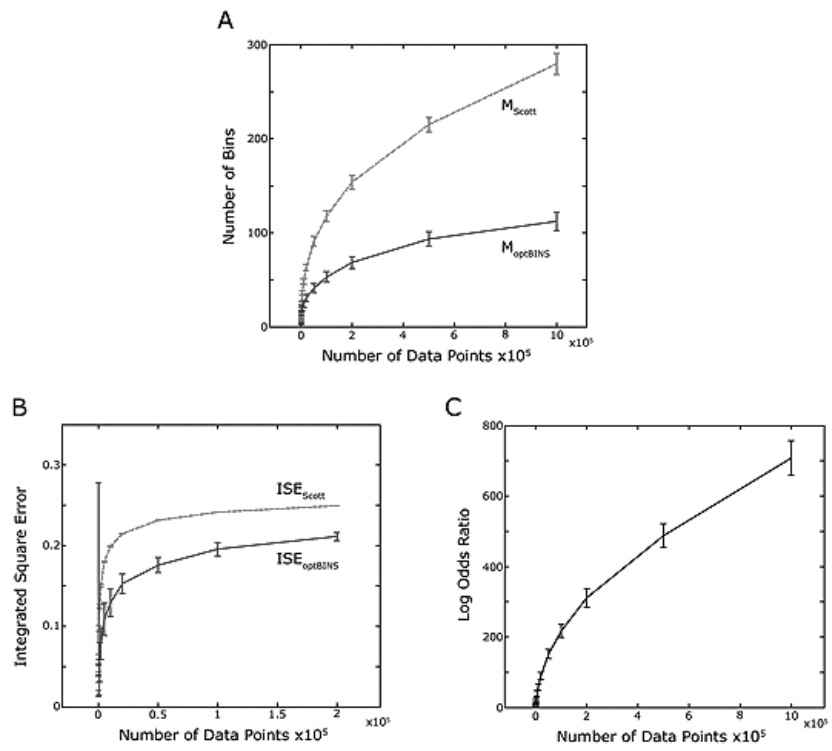


Fig. 2.— Describe these results

## 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

$$\begin{aligned} p(M = 1|d_1, d_2, I) &\propto M^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})} \\ &\propto 1^2 \frac{\Gamma(\frac{1}{2})}{\Gamma(\frac{1}{2})^1} \frac{\Gamma(2 + \frac{1}{2})}{\Gamma(2 + \frac{1}{2})} = 1, \end{aligned} \quad (55)$$

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

$$\begin{aligned} p(M|d_1, d_2, I) &\propto M^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{k=1}^M \Gamma(n_k + \frac{1}{2})}{\Gamma(N + \frac{M}{2})} \\ &\propto M^2 \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\Gamma(1 + \frac{1}{2})^2 \Gamma(\frac{1}{2})^{M-2}}{\Gamma(2 + \frac{M}{2})} \\ &\propto M^2 \frac{\Gamma(\frac{3}{2})^2}{\Gamma(\frac{1}{2})^2} \frac{\Gamma(\frac{M}{2})}{\Gamma(2 + \frac{M}{2})} \\ &\propto \frac{1}{2} \cdot \frac{M}{1 + \frac{M}{2}}. \end{aligned} \quad (56)$$

Figure 3A 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

$$p(M|d_1, d_2, d_3, I) \propto \frac{3}{4} \cdot \frac{M^2}{(2 + \frac{M}{2})(1 + \frac{M}{2})}, \quad (57)$$

and for each point in a separate bin we have

$$p(M|d_1, d_2, d_3, I) \propto \frac{1}{4} \cdot \frac{M^2}{(2 + \frac{M}{2})(1 + \frac{M}{2})}. \quad (58)$$

While the logarithm of the posterior in (57) can be greater than zero, as  $M$  increases, the data points eventually fall into separate bins. This causes the posterior to change from (57) to (58) resulting in a dramatic decrease in the logarithm of the posterior, which then asymptotically increases to zero as  $M \rightarrow \infty$ .

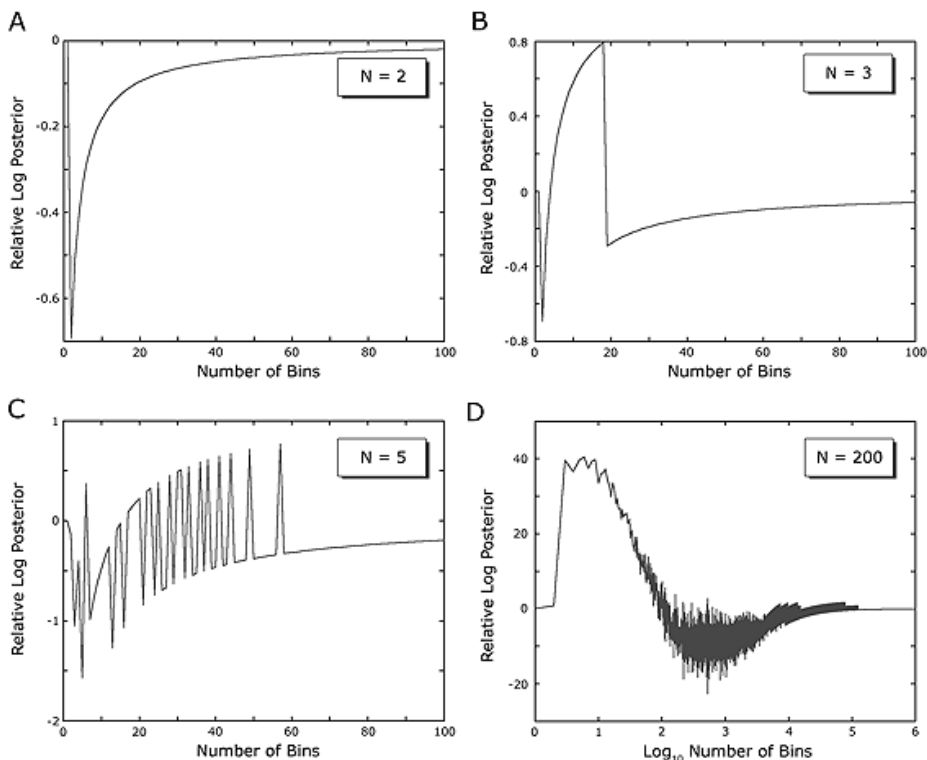


Fig. 3.— Describe these results

This behavior is shown in Figure 3B. The result is that either one or two bins will be optimal depending on the relative positions of the data points.

More rich behavior can be seen in the case of  $N = 5$  data points. The results again (Figure 3C) 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 optimal number of bins depends sensitively on the relative positions of the samples.

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

$$p(M|\underline{d}, I) \propto \left(\frac{M}{2}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(N + \frac{M}{2})}, \quad (59)$$

which again results in a log posterior that asymptotically approaches zero as  $M \rightarrow \infty$ . Figure 3D

demonstrates these two effects for  $N = 200$ . This can be compared to the log posterior for 1000 Gaussian samples in Figure 1B.

## 5.2. Sufficient Data

This investigation on the effects of small sample size 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 (eg. Fig. 1B, Fig. 3D). With small sample sizes, however, one finds that the bin heights have large error bars (Figure 4A) so that  $\mu_i \simeq \sigma_i$ , and that the log posterior is multi-modal (Figure 4B) 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 of  $N$  was used as an indicator of the stability of the solution.

Figure 4C 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 100 samples, and preferentially 150 samples, are required to accurately and consistently infer the shape of the 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 Bayman and Broadhurst(1979). 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 the ability of the optBINS algorithm to detect severe losses of information due to rounding or truncation.

In the event that data have been severely rounded, there is the possibility of multiple identical data points. A severe amount of information has been lost when this is a more prevalent feature than the general shape of the probability density.

\*\*\*Describe the asymptotic solution as  $M \rightarrow \infty$

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  $n_i$  data points in the  $i^{th}$  bin and the posterior probability can be written as

$$p(M|\underline{d}, I) \propto \left(\frac{M}{2}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(N + \frac{M}{2})} \cdot \prod_{i=1}^{M-E} (2n_i - 1)!!, \quad (60)$$

where  $E$  is the number of empty bins and  $i$  indexes the  $M - E$  non-empty bins. For  $M > M_{crit}$ , as  $M \rightarrow \infty$ , the log posterior asymptotes to  $\sum_{i=1}^{M-E} \log((2n_i - 1)!!)$ , which can be further simplified to

$$\sum_{i=1}^{M-E} \log((2n_i - 1)!!) = (M - E - N) \log(2) + \sum_{i=1}^{M-E} \sum_{s=n_i}^{2n_i-1} \log s. \quad (61)$$

To test for excessive rounding or truncation,  $\log p(M|\underline{d}, I)$  for  $M < M_{crit}$  should be compared to (61) above. If the latter is larger, than the discrete nature of the data is a more significant effect than the general shape of the underlying probability density function. A reasonable histogram can still be obtained by adding a uniformly-distributed random number with a range defined by the discretization to each datum point Bayman and Broadhurst(1979). While this will produce the best histogram possible given the data, this will not recover the lost information.

## 7. 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-pieceswise

model  $h(x, y)$  of the two-dimensional density function  $f(x, y)$  is

$$h(x, y; M_x, M_y) = \frac{M}{V} \sum_{j=1}^{M_x} \sum_{k=1}^{M_y} \pi_{j,k} \Pi(x_{j-1}, x, x_j) \Pi(y_{k-1}, y, y_k), \quad (62)$$

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  $M - 1$  model 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

$$p(d_n | \pi_{j,k}, M_x, M_y, I) = \frac{M}{V} \pi_{j,k}. \quad (63)$$

The previous prior assignments result in the posterior probability

$$p(\underline{x}, M_x, M_y | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \prod_{j=1}^{M_x} \prod_{k=1}^{M_y} \pi_{j,k}^{n_{j,k} - \frac{1}{2}}, \quad (64)$$

where  $\pi_{M_x, M_y}$  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

$$p(M_x, M_y | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{j=1}^{M_x} \prod_{k=1}^{M_y} \Gamma(n_{j,k} + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \quad (65)$$

where  $M = M_x M_y$ .

For a D-dimensional histogram, the general result is

$$p(M_1, \dots, M_D | \underline{d}, I) \propto \left(\frac{M}{V}\right)^N \frac{\Gamma(\frac{M}{2})}{\Gamma(\frac{1}{2})^M} \frac{\prod_{i_1=1}^{M_1} \dots \prod_{i_D=1}^{M_D} \Gamma(n_{i_1, \dots, i_D} + \frac{1}{2})}{\Gamma(N + \frac{M}{2})}, \quad (66)$$

where  $M_i$  is the number of bins along the  $i^{\text{th}}$  dimension,  $M$  is the total number of bins,  $V$  is the D-dimensional volume of the histogram, and  $n_{i_1, \dots, i_D}$  indicates the number of counts in the bin indexed by the coordinates  $(i_1, \dots, i_D)$ . Note that the result in (36) can be used directly for a multi-dimensional histogram simply by relabelling the multi-dimensional bins with a single index.

Figure 5 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. Figure 5A shows the relative logarithm of the posterior probability plotted as a function of the number of bins in each dimension. The same surface is displayed as contour plot in Figure 5B, where we find the optimal number of bins to be  $12 \times 14$ . Figure 5C shows the optimal two-dimensional histogram model. Note that modelled density function is displayed in terms of the number of counts rather than the probability density, which can be easily computed using (49) with error bars computed using (51). In Figure 5D, we show the histogram obtained using Stone’s method, which results in a  $27 \times 28$  array of bins. This is clearly a sub-optimal model since random sampling variations are easily visible.

## 8. Discussion

Discuss interplay between prior and likelihood: Occam factor

There is a source of error not considered in this analysis: the uncertainty in the number of bins.

### Appendix 1: Matlab code for the optimal number of uniform bins in a histogram

```
% optBINS computes the optimal number of bins for a given one-dimensional
% data set. This optimization is based on the posterior probability for
% the number of bins
%
% Usage:
%       optM = optBINS(data,minM,maxM);
%
% Where:
%       data is a (1,N) vector of data points
%       minM is the minimum number of bins to consider
%       maxM is the maximum number of bins to consider
%
% 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.
%
% Created by Kevin H. Knuth on 17 April 2003
% Modified by Kevin H. Knuth on 21 February 2006

function optM = optBINS(data,minM,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 = minM:maxM
    n = hist(data,M); % Bin the data (equal width bins here)
    logp(M) = N*log(M) + gammaln(M/2) - gammaln(N+M/2) - M*gammaln(1/2) + sum(gammaln(n+0.5));
end [maximum, optM] = max(logp(minM,maxM)); return
```

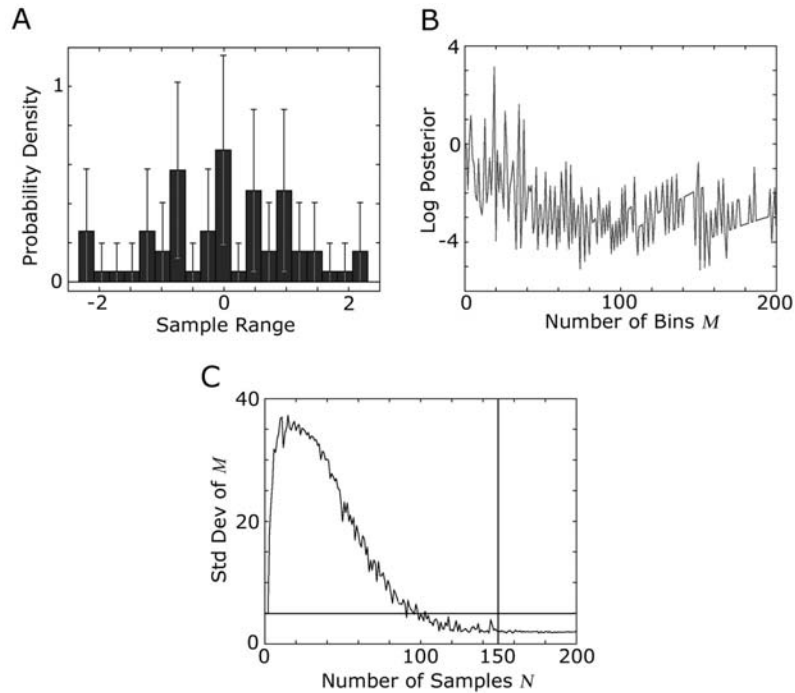


Fig. 4.— (A) An optimal density model ( $M = 19$ ) for  $N = 30$  data points sampled from a Gaussian distribution. The fact that the error bars on the bin probabilities are as large as the probabilities themselves indicates that this is a poor estimate. (B) The log posterior probability for the number 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.

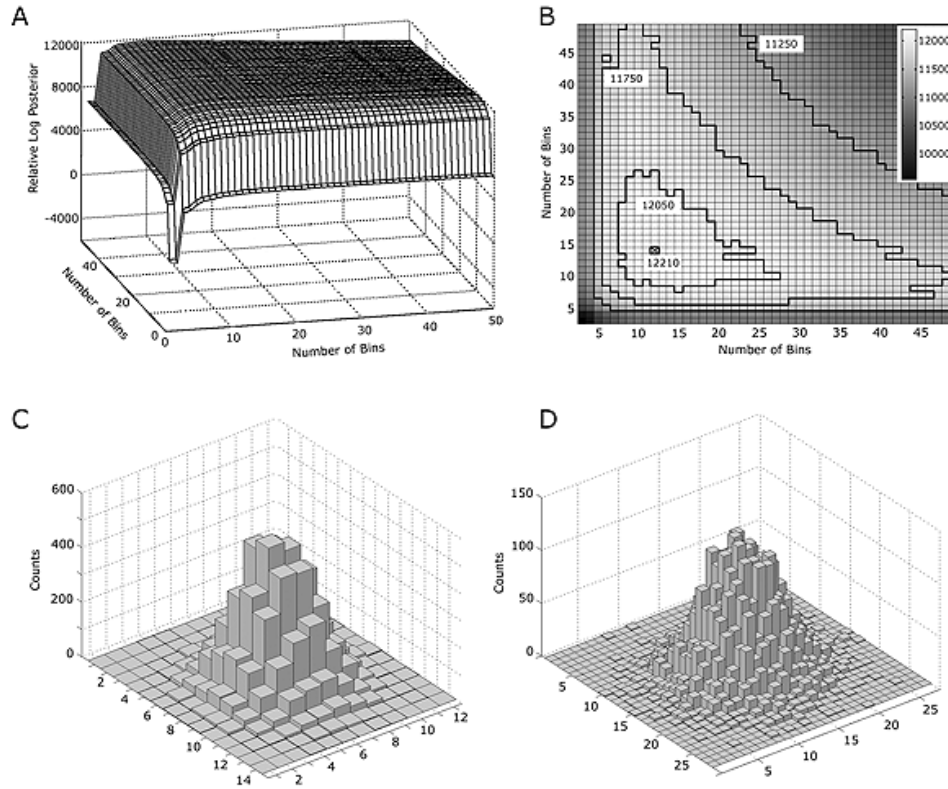


Fig. 5.— 10000 samples were drawn from a two-dimensional Gaussian density to demonstrate the optimization of a two-dimensional histogram. (A) The relative logarithm of the posterior probability is plotted as a function of the number of bins in each dimension. The normalization constant has been neglected 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  $12 \times 14$ . (C) The optimal histogram for this data set. (D) The histogram determined using Stone’s method has  $27 \times 28$  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.

## REFERENCES

- Abramowitz, M. and Stegun, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover Publications, Inc., p. 255.
- Akaike, H. (1974). A new look at the statistical identification model. *IEEE Trans. Automatic Control*. **19**, 716–723.
- Bayman, B. F. and Broadhurst, J. B. (1979). A simple solution to a problem arising from the processing of finite accuracy digital data using integer arithmetic. *Nuclear Instruments and Methods*. **167**, 475–478.
- Berger, J. O. and Bernardo, J. M. (1992). Ordered group reference priors with application to the multinomial problem. *Biometrika*. **79**, 25–37.
- Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis*. New York: John Wiley & Sons, p. 55.
- Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator:  $L_2$  theory. *Zeitschrift für Wahrscheinlichkeitstheorie verw. Gebiete*. **57**, 453–476.
- Hartigan, J. A. (1996). Bayesian histograms. *Bayesian Statistics* (eds. J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith) vol. 5, pp. 211–222. Oxford: Oxford Univ. Press.
- Jeffreys, H. (1961). *Theory of Probability* 3rd. ed. Oxford: Oxford University Press.
- Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*. **9**, 65–78.
- Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*. **66**, 605–610.
- Stone, C. J. (1984). An asymptotically histogram selection rule. *Proc. Second Berkeley Symp* (ed. J. Neyman) pp. 513–520. Berkeley: Univ. California Press.