On Selecting The Number Of Bins For A Histogram

Sai Venu Gopal Lolla, Lawrence L. Hoberock
School of Mechanical and Aerospace Engineering
Oklahoma State University, Stillwater OK 74078

Abstract— *Histograms are widely used in exploratory data analysis for graphically describing datasets. This paper presents a new method for selecting the number of bins to be used for constructing a histogram for a given dataset. The improved performance of the proposed method is compared to the performances of methods proposed by Sturges, Scott, Freedman et al., Shimazaki et al., and Knuth.*

Keywords: histogram; bin selection;

1. Introduction

A histogram is a graphical representation of the frequency distribution of a dataset. Widely employed in exploratory data analysis, a histogram can be treated as a simple non–parametric density estimator. For a given dataset, a histogram can visually convey the information relating to shape, spread, location, modality and symmetry of the distribution of the underlying population, and is well suited for summarizing large datasets [10]. While more sophisticated kernel–based density estimators are available, histograms are widely employed due to the ease and simplicity of construction and interpretation [20], [16]. While histograms are used mainly for visualizing data and obtaining summary quantities such as entropy, the values of such quantities depend upon the number of bins used (or the bin width used) and the location of the bins [7].

Let $X = \{x_1, x_2, \ldots, x_n\}$ be a univariate dataset with probability density function $f(x)$. We follow Martinez et al. [10]: To construct a histogram, an origin for the bins t_0 (also referred to as the anchor) and a bin width h are selected. Selection of these two parameters defines a mesh (position of all the bins) over which the histogram will be constructed. Each bin is represented by a pair of bin edges as $B_k = [t_k, t_{k+1})$, where $t_{k+1} - t_k = h$ for all k. Histograms using varying bin widths are not addressed in this paper. Let c_k represent the number of observations in B_k (bin count for B_k) given by:

$$c_k = \sum_{i=1}^{n} I_{B_k}(x_i) \tag{1}$$

where I_{B_k} is defined as:

$$I_{B_k}(x_i) = \begin{cases} 1 & x_i \text{ in } B_k \\ 0 & x_i \text{ not in } B_k \end{cases} \tag{2}$$

While the density estimate for the underlying population (c_k for all k) satisfies the non–negativity condition necessary for it to be a *bona fide* probability density function, the summation of all the probabilities do not necessarily add to unity. To satisfy that condition, the probability density function estimate, $\hat{f}(x)$, as obtained from a histogram, is defined as:

$$\hat{f}(x) = \frac{c_k}{nh} \qquad \text{for } x \text{ in } B_k \tag{3}$$

This assures that $\int \hat{f}(x)dx = 1$ is satisfied, and $\hat{f}(x)$ represents a valid estimate for the probability density function of the population underlying the dataset.

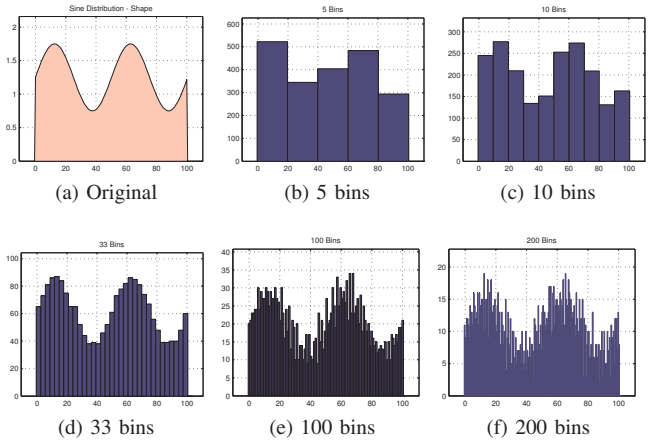
The information relating to shape, modality, symmetry and summary quantities estimated using a histogram will depend on the values that c_k (and $\hat{f}(x)$) assume, which in turn depend upon the parameters t_0 and h.

While histograms are commonly constructed using $t_0 = \min(X)$, it is known that modifying this parameter can sometimes cause a rather drastic change in the values assumed by c_k [21]. Simonoff et al. [16] provide a method to quantify the effects of changing the parameter t_0 during the construction of a histogram. However, in the work herein, we use $t_0 = \min(X)$.

A common method to determine bin width h is:

$$h = \frac{\max(X) - \min(X)}{m} \tag{4}$$

where m is the number of bins. From (1), (2), (3) and (4) it can be seen that the number of bins used to construct a histogram will influence c_k (and $\hat{f}(x)$) and any further information derived from them. Consider the following two extreme cases: (1) Using only one bin ($m = 1$) will cause all the data points in X to map to that bin, and information



(a) Original (b) 5 bins (c) 10 bins

(d) 33 bins (e) 100 bins (f) 200 bins

Fig. 1
ORIGINAL DISTRIBUTION AND SEVERAL HISTOGRAMS FOR A DATASET (≈ 2000 POINTS)

Datafile	# of Datasets	# of Data points
DF–1	12	≈ 500
DF–2	12	≈ 1000
DF–3	12	≈ 2000
DF–4	12	≈ 5000

Table 1
DATAFILES USED FOR TESTING

Dataset	Distribution	Dataset	Distribution
DS–1	Uniform	DS–7	Gamma
DS–2	Sine	DS–8	Triangular
DS–3	Normal	DS–9	Custom-1
DS–4	Laplace	DS–10	Custom-2
DS–5	Semi-Circular	DS–11	Custom-3
DS–6	Exponential	DS–12	Custom-4

Table 2
DATAFILES USED FOR TESTING

relating to shape, modality, and symmetry will be lost (unless the underlying population distribution is Uniform); (2) Using n or more bins ($m \geq n$) will spread the data points over all the bins more or less uniformly, such that any information relating to shape, modality, and symmetry will again be lost. These two extreme cases suggest that an "optimal" number of bins should be used to construct a histogram that can effectively capture information relating to shape, modality, and symmetry and provide meaningful values for summary quantities. Using very few bins (small value for m) results in a large bin width, and hence a histogram that captures the shape of the underlying distribution "coarsely" (under–fitting). Using excessive bins (large value for m) results in a small bin width, and hence a "noisy" histogram that captures the shape of the underlying distribution "finely" and typically "noisily" (over–fitting). *Fig.1 illustrates that arbitrarily increasing the number of bins to construct a histogram does not necessarily result in "better" histograms.*

Thus, the problem of selecting an "optimal" number of bins refers to selecting an appropriate number of bins for constructing a histogram that achieves a "good" balance between "degree of detail" and "noisiness" for a given dataset. In other words, the number of bins should be large enough to capture all the major shape features present in the distribution, but small enough so as to suppress finer details produced due to random sampling noise [7].

Tables 1 and 2 and Fig.2 describe the datafiles and datasets used for testing our proposed method.

2. Existing Methods

Perhaps the earliest reported method for constructing histograms is due to Sturges [18]. It is based on the assumption that a good distribution will have binomial coefficients $\binom{m-1}{i}, i = 0, 1, 2, \ldots, m-1$ as its bin counts. It suggests the number of bins to be used as:

$$m = 1 + \log_2 n \tag{5}$$

Hyndman [6] suggests that the argument used by Sturges [18] is incorrect and should not be used. Scott [14] uses IMSE (Integrated Mean Square Error – which is equal to Mean Integrated Square Error MISE [11]) as the measure of error between the estimated probability density ($\hat{f}(x)$) represented by the histogram, and the actual (and unknown) probability density ($f(x)$) of the underlying population. MISE is defined as:

$$\begin{aligned} IMSE &= \int MSE(x)\, dx \\ &= \int E(\hat{f}(x) - f(x))^2\, dx \\ &= E \int (\hat{f}(x) - f(x))^2\, dx \\ &= MISE \end{aligned} \tag{6}$$

Using this error metric with Gaussian density as the reference for the actual probability density, Scott suggests a bin width of:

$$h = \frac{3.49s}{n^{1/3}} \tag{7}$$

where s is the estimated standard deviation. Freedman et al. [2] suggests a similar formula with a slight modification as:

$$h = \frac{2(IQR(X))}{n^{1/3}} \tag{8}$$

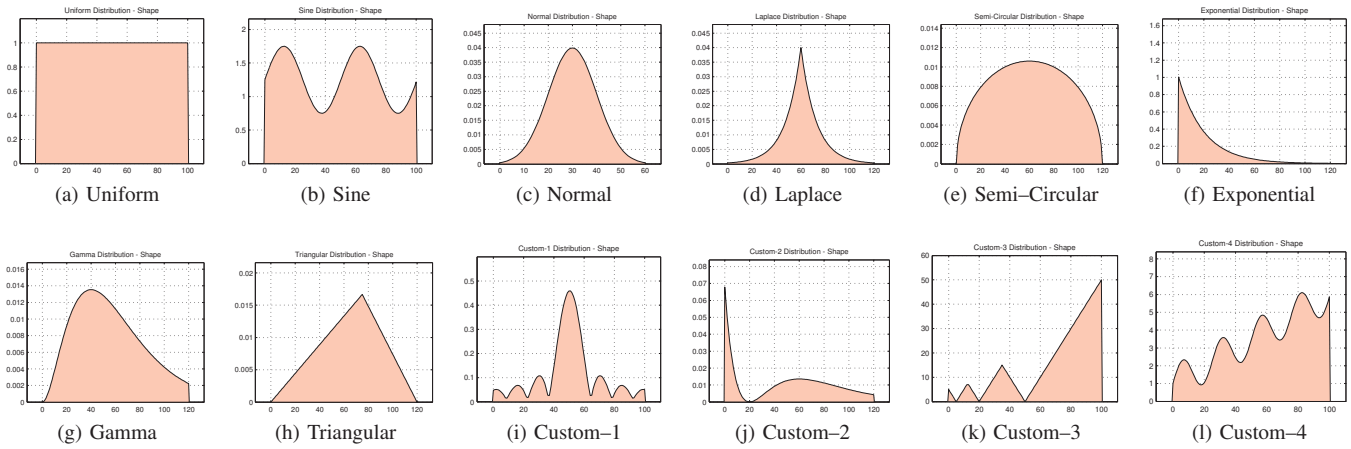
where $IQR(X)$ is the Inter–Quartile Range for the dataset X.

Methods proposed by Stone [17], Rudemo [12], and Wand [20] are also frequently encountered in the related literature. Stone [17] proposes a method based on minimization of a loss function defined on the basis of bin probabilities and number of bins. Rudemo [12] proposes a method based on Kullback–Leibler risk function and cross–validation techniques. Wand [20] extends Scott's method [14] to have good large sample consistency properties. Hall [5] investigates the use of Akaike's Information Criterion (AIC) and Kullback Liebler Cross Validation methods for constructing histograms.

More recently, Birge et al. [8] have proposed a method using a risk function based on penalized maximum likelihood. Knuth [7] has proposed a method based on maximizing the posterior probability for number of bins. Shimazaki et al. [15] have proposed a method based on minimizing an estimated cost function obtained by using a modified MISE. The method evaluates the estimated cost function using the implications of an assumption that the data are sampled independently of each other (assumption of a Poisson point process).

3. A New Proposed Method

Popular methods such as those given by Scott [14], and Freedman et al. [2] try to asymptotically minimize MISE. These methods make certain assumptions to allow estimating the value of MISE, since the actual density function of the underlying population itself is unknown. Knuth [7] suggests that it is not reasonable to extend these assumptions for all datasets. It is also known that MISE does not necessarily conform with the human perception of closeness of a density function to its target [21]. Marron et al. [9] provide a good introduction to the disconnect between classical mathematical theory and the practice of non–parametric density estimation due to the non–conformance of human perception of closeness with metrics such as MISE and MIAE. Methods employing risk functions based on penalized likelihood functions need not make assumptions about the underlying function, but their



(a) Uniform (b) Sine (c) Normal (d) Laplace (e) Semi–Circular (f) Exponential

(g) Gamma (h) Triangular (i) Custom–1 (j) Custom–2 (k) Custom–3 (l) Custom–4

Fig. 2

DATASETS USED FOR TESTING

performance will depend upon the form of the risk function selected.

In the new method proposed here, error metrics are defined on quantities observable or computable from the dataset. An intuitive balance between the error and the cost of computing the histogram is used to select the number of bins.

Motivation: A histogram for a given dataset can be interpreted as a compact representation of the dataset itself, obtained by a lossy compression process. A good histogram will provide enough information to recreate data whose Cumulative Distribution Function (CDF) approximately matches the Cumulative Distribution Function of the actual dataset itself (Statement–I). Also, a good histogram will have no significant shape information inside any bin (Statement–II).

Reflection will show that Statements I & II are axiomatic. They also indicate that data can be reconstructed from a given histogram. There are two simple ways to approximately reconstruct data from a histogram. For each bin B_k with bin count c_k: (1) recreate c_k data points equal to the bin center $((t_k+t_{k+1})/2)$ – equivalent to nearest neighbor interpolation; (2) recreate c_k data points spread uniformly over (t_k, t_{k+1}) – equivalent to linear interpolation.

Let $\hat{X}_{NN} = \{\hat{x}_{1_{NN}}, \hat{x}_{2_{NN}}, \ldots, \hat{x}_{n_{NN}}\}$ represent data reconstructed using the nearest neighbor equivalent described above, and let $\hat{X}_L = \{\hat{x}_{1_L}, \hat{x}_{2_L}, \ldots, \hat{x}_{n_L}\}$ represent data reconstructed using the linear interpolation equivalent. Fig.3 illustrates that for a histogram constructed using a given number of bins for a dataset, the CDF of the data recreated using linear interpolation matches the actual CDF more closely than the data recreated using the nearest neighbor approximation. Due to the Glivenko-Cantelli theorem [3], [1] both approximations will converge to the actual CDF itself as m increases.

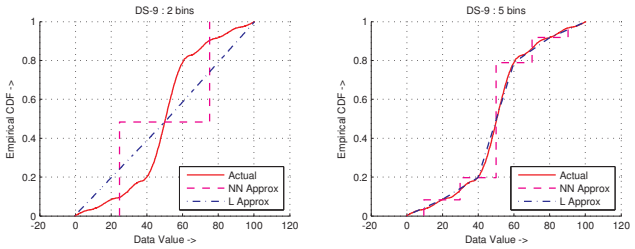
Define the error metrics E_{NN} and E_L for the nearest neighbor and linear interpolation reconstructions , respectively, by:

$$\begin{aligned} E_{NN} &= \sum_{i=1}^{n} |x_i - \hat{x}_{i_{NN}}| \\ E_L &= \sum_{i=1}^{n} |x_i - \hat{x}_{i_L}| \end{aligned} \tag{9}$$

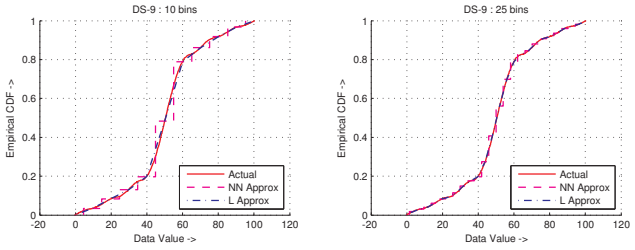
Due to the aforementioned theorem, E_{NN} and E_L will converge to zero as the number of bins used to construct the histogram are increased ($m \to \infty$). In fact the convergence of the error metrics to zero is very likely once $m \geq n$. The CDF of data reconstructed using linear interpolation, which matches the actual data CDF more closely than the data reconstructed using the nearest neighbor approximation, indicates that E_L will converge faster than E_{NN}. Fig.4 shows plots of E_{NN} and E_L for various values of m. In these plots, the vertical axis represents the value of the error metrics, and the horizontal axis represents the value of the computational cost. The computational cost involved in constructing a histogram using m bins for n points will at the most be of order $O(mn)$. Since we are trying to select m for the same n points, the computational costs will be proportional to m and hence m is used as the computational cost.

Fig.4 uses square markers to indicate "elbow points" for both error metric curves. An elbow point marks the region where incurrence of further "costs" does not result in further significant "gains". Hence elbow points represent an intuitive trade–off between two conflicting quantities. The method is often traced to Thorndike [19] and has been used for similar purposes [22], [13]. The method described in [22] is used to compute the elbow points for the work done in this paper.

Let m_{NN} and m_L correspond, respectively, to the number of bins indicated by the elbow points on the E_{NN} and E_L metric curves. Using any m in $[m_L,m_{NN}]$ will result in a histogram that offers a reasonably good trade–off between the error metrics and the cost involved. In all the histograms constructed using an m in $[m_L,m_{NN}]$, the histogram having



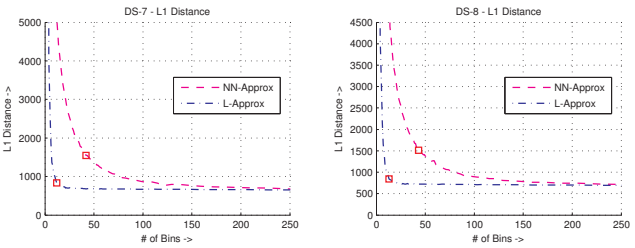
(a) $m = 2$ (b) $m = 5$



(c) $m = 10$ (d) $m = 25$

Fig. 3

EMPIRICAL CDF: DATA APPROXIMATIONS USING $m = 2, 5, 10, 25$ BINS FOR DS–9 (DF–3)



(a) DS–7 (b) DS–8

Fig. 4

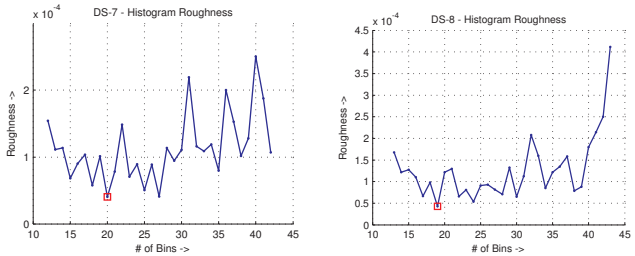
ERROR METRICS FOR DS–7 & DS–8 (DF–3)

the lowest roughness $\hat{R}$ is likely to be the most visually appealing. The roughness measure for a histogram is defined as [4]:

$$\hat{R} = \sum(\Delta^2 \hat{f}(x))h \tag{10}$$

where Δ^2 represents the second order finite difference for $\hat{f}(x)$. Fig.5 shows Roughness measures for histograms constructed with m in the corresponding $[m_L, m_{NN}]$ for DS–7 & DS–8.

In summary, to construct a histogram using our new method: (1) Define $M_1 = \{1, 2, \ldots, \sqrt{n}, \frac{n}{\sqrt{n}}, \ldots, \frac{n}{2}, \frac{n}{1}\}$; (2) Construct a histogram for X with m bins for all m in M_1; (3) Construct E_{NN} and E_L for each histogram; (4) Compute m_{NN} and m_L for the E_{NN} and E_L metric curves; (5) Define $M_2 = \{m_L, m_L + 1, \ldots, m_{NN} - 1, m_{NN}\}$; (6) For each m in M_2 construct a histogram for X with m bins; (7) Compute roughness metric $\hat{R}$ for each histogram; (8) Select as the optimal number of bins m_{opt}, the value of m that has the lowest $\hat{R}$.



(a) DS–7 (b) DS–8

Fig. 5

ROUGHNESS MEASURES FOR DS–7 & DS–8 (DF–3)

4. Experiments & Results

The method explained in Section 3 was coded in MATLAB for testing on the datafiles/datasets introduced in Section 1. Shimazaki et al. [15] and Knuth [7] provide MATLAB implementations of their methods. Methods due to Sturges [18], Scott [14], and Freedman et al. [2] were also coded in MATLAB. All the methods were tested on datafiles DF–1, DF–2, DF–3, and DF–4.

The following abbreviations are used in the tables and figures displaying results: StM – Sturges Method; ScM – Scott Method; FDM – Freedman Diaconis Method; SM – Shimazaki et al. Method; KM – Knuth Method; LHM – method proposed in this paper.

In order to measure the performance of the various methods mentioned above, the values of E_{NN}, E_L, and $\hat{R}$ are computed for the histograms generated by each method. It is desirable to have values as low as possible for all three metrics simultaneously. However, low values of $\hat{R}$ tend to result in relatively higher values of E_{NN} and E_L, and vice versa. E_{NN} and E_L indicate a given histogram's fidelity in representing the data, and $\hat{R}$ indicates the degree of over–fitting (or under–fitting) in the representation.

Tables 3 and 4 document values of m_{opt}, E_{NN}, E_L, and $\hat{R}$ for histograms generated by various methods for each dataset. The maximum values for m_{opt}, and the minimum values for E_{NN}, E_L, and $\hat{R}$ across all the methods are highlighted in blue boldface for easy reading. It can be seen from the tables that the method proposed herein (LHM) produces the lowest values of E_{NN}, E_L, and $\hat{R}$ simultaneously for a vast majority of the cases. This indicates that the proposed method does a better job of capturing shape-related information to a good degree of detail without admitting excessive noise as compared to the other methods.

Fig.6 to 11 display histograms constructed using various methods for the datasets in datafile DF–3. Visual examination of these plots and comparison to data distribution shapes in Fig.2 supports the aforementioned inference. Results for other datasets (and other datafiles) were found to be similar.

The method proposed in this paper also produces some results that the authors find less satisfying, in which case the shapes of the distributions underlying the population are not as well captured. However, as shown in Fig.12 and 13, results from the other methods are also less satisfying.

DS		StM	ScM	FDM	SM	KM	LHM
DS-1	m_{opt}	10	**13**	**13**	1	1	1
	E_{NN} (x 10^2)	12.75	**9.86**	**9.86**	127.46	127.46	127.46
	E_L (x 10^2)	1.84	**1.75**	**1.75**	1.90	1.90	1.90
	R (x 10^{-5})	2.19	16.95	16.95	**0.00**	**0.00**	**0.00**
DS-2	m_{opt}	11	12	12	4	4	**22**
	E_{NN} (x 10^2)	12.68	11.62	11.62	34.81	34.81	**6.43**
	E_L (x 10^2)	2.12	2.18	2.18	2.59	2.59	**1.87**
	R (x 10^{-4})	6.13	4.92	4.92	37.06	37.06	**2.20**
DS-3	m_{opt}	11	5	3	8	8	**20**
	E_{NN} (x 10^2)	7.39	15.91	26.28	10.04	10.04	**4.04**
	E_L (x 10^2)	1.97	4.81	9.00	2.30	2.30	**1.73**
	R (x 10^{-4})	18.00	142.05	499.85	43.61	43.61	**2.25**
DS-4	m_{opt}	11	8	5	15	11	**28**
	E_{NN} (x 10^2)	15.18	21.87	32.05	11.44	15.18	**6.17**
	E_L (x 10^2)	3.46	7.71	9.43	2.50	3.46	**2.03**
	R (x 10^{-3})	9.22	9.80	33.42	5.26	9.22	**1.05**
DS-5	m_{opt}	11	13	12	6	3	**15**
	E_{NN} (x 10^3)	1.53	1.27	1.40	2.77	5.43	**1.12**
	E_L (x 10^2)	2.36	2.07	**2.05**	3.14	9.08	2.10
	R (x 10^{-5})	14.59	12.91	19.88	19.04	166.64	**8.16**
DS-6	m_{opt}	11	11	7	14	7	**18**
	E_{NN} (x 10^2)	15.79	15.79	25.09	12.51	25.09	**9.87**
	E_L (x 10^2)	3.26	3.26	6.20	2.46	6.20	**2.33**
	R (x 10^{-4})	7.08	7.08	23.04	3.57	23.04	**2.03**
DS-7	m_{opt}	11	**12**	10	7	7	**12**
	E_{NN} (x 10^3)	1.56	**1.39**	1.67	2.41	2.41	**1.39**
	E_L (x 10^2)	2.77	**2.51**	2.62	3.65	3.65	**2.51**
	R (x 10^{-4})	1.67	**1.19**	2.69	8.47	8.47	**1.19**
DS-8	m_{opt}	11	11	9	9	6	**19**
	E_{NN} (x 10^2)	15.41	15.41	18.65	18.65	27.79	**8.92**
	E_L (x 10^2)	2.49	2.49	2.95	2.95	5.12	**2.13**
	R (x 10^{-4})	2.38	2.38	4.70	4.70	13.78	**1.17**
DS-9	m_{opt}	11	9	4	19	5	**27**
	E_{NN} (x 10^2)	12.61	15.43	40.00	7.30	25.34	**5.43**
	E_L (x 10^2)	4.47	5.05	20.38	2.35	5.54	**2.15**
	R (x 10^{-3})	12.56	27.60	6.77	5.83	51.06	**2.32**
DS-10	m_{opt}	11	15	17	**24**	16	23
	E_{NN} (x 10^2)	15.84	11.48	10.18	**7.29**	11.03	7.54
	E_L (x 10^2)	4.37	3.24	2.67	2.25	2.82	**2.17**
	R (x 10^{-4})	52.63	30.38	22.56	11.38	24.56	**8.65**
DS-11	m_{opt}	11	11	8	8	7	**20**
	E_{NN} (x 10^2)	12.57	12.57	17.50	17.50	20.02	**6.93**
	E_L (x 10^2)	2.75	2.75	3.76	3.76	4.20	**2.12**
	R (x 10^{-4})	15.65	15.65	19.44	19.44	36.89	**4.33**
DS-12	m_{opt}	11	**12**	11	4	4	4
	E_{NN} (x 10^3)	1.26	**1.16**	1.26	3.46	3.46	3.46
	E_L (x 10^2)	**2.25**	2.32	**2.25**	3.06	3.06	3.06
	R (x 10^{-6})	2036.87	1678.39	2036.87	**9.75**	**9.75**	**9.75**

(a) Results for DF–1 ($\approx$ 500 points)

DS		StM	ScM	FDM	SM	KM	LHM
DS-1	m_{opt}	**11**	10	10	1	1	1
	E_{NN} (x 10^3)	**2.34**	2.57	2.57	25.50	25.50	25.50
	E_L (x 10^2)	3.41	3.48	3.48	**3.41**	**3.41**	**3.41**
	R (x 10^{-5})	2.58	5.25	5.25	**0.00**	**0.00**	**0.00**
DS-2	m_{opt}	12	10	10	4	4	**28**
	E_{NN} (x 10^2)	22.23	10	26.39	26.39	66.19	**9.76**
	E_L (x 10^2)	3.97	26.39	4.60	4.60	5.72	**3.54**
	R (x 10^{-4})	6.59	4.60	10.03	10.03	41.49	**1.13**
DS-3	m_{opt}	12	4	3	14	10	**18**
	E_{NN} (x 10^2)	13.41	38.86	51.60	11.44	15.80	**9.13**
	E_L (x 10^2)	4.06	16.06	19.13	3.70	4.27	**3.62**
	R (x 10^{-4})	14.10	169.16	514.27	6.76	19.19	**2.94**
DS-4	m_{opt}	12	6	4	21	11	**37**
	E_{NN} (x 10^2)	27.22	56.19	88.13	15.52	28.77	**9.06**
	E_L (x 10^2)	7.22	25.40	55.74	3.97	6.59	**3.75**
	R (x 10^{-4})	55.57	156.80	103.28	25.63	99.27	**5.61**
DS-5	m_{opt}	12	10	10	6	6	**18**
	E_{NN} (x 10^3)	2.65	3.11	3.11	5.16	5.16	**1.75**
	E_L (x 10^2)	4.03	4.65	4.65	6.33	6.33	**3.68**
	R (x 10^{-5})	5.02	5.56	5.56	22.77	22.77	**4.07**
DS-6	m_{opt}	12	8	5	18	9	**28**
	E_{NN} (x 10^3)	2.69	4.09	6.66	1.81	3.61	**1.17**
	E_L (x 10^2)	5.36	10.56	24.50	**3.75**	7.59	3.83
	R (x 10^{-4})	6.01	16.52	50.55	3.25	12.33	**1.00**
DS-7	m_{opt}	12	9	8	13	9	**19**
	E_{NN} (x 10^3)	2.66	3.54	3.96	2.47	3.54	**1.69**
	E_L (x 10^2)	4.78	5.60	6.53	4.31	5.60	**3.82**
	R (x 10^{-5})	12.98	36.91	59.32	12.62	36.91	**4.39**
DS-8	m_{opt}	12	9	7	10	10	**19**
	E_{NN} (x 10^3)	2.63	3.49	4.46	3.15	3.15	**1.67**
	E_L (x 10^2)	4.21	5.68	8.19	4.96	4.96	**3.62**
	R (x 10^{-5})	20.62	42.13	81.31	28.39	28.39	**4.80**
DS-9	m_{opt}	12	7	3	30	18	**36**
	E_{NN} (x 10^2)	22.30	41.00	69.76	9.32	14.99	**7.76**
	E_L (x 10^2)	6.92	8.41	29.66	3.84	4.67	**3.77**
	R (x 10^{-3})	17.98	42.55	30.02	2.04	7.68	**1.27**
DS-10	m_{opt}	12	12	14	**47**	17	30
	E_{NN} (x 10^2)	27.27	27.27	23.49	**7.26**	19.30	11.01
	E_L (x 10^2)	7.98	7.98	6.30	**3.59**	5.05	3.80
	R (x 10^{-4})	53.64	53.64	40.98	6.88	24.53	**5.71**
DS-11	m_{opt}	12	9	6	16	13	**25**
	E_{NN} (x 10^3)	2.23	3.00	4.46	1.66	2.08	**1.10**
	E_L (x 10^2)	4.85	6.83	13.73	4.19	4.43	**3.66**
	R (x 10^{-4})	14.30	14.98	16.36	7.15	12.02	**3.77**
DS-12	m_{opt}	12	9	9	8	4	**24**
	E_{NN} (x 10^3)	2.22	2.95	2.95	3.32	6.61	**1.15**
	E_L (x 10^2)	4.52	5.59	5.59	4.31	6.68	**3.77**
	R (x 10^{-6})	1949.62	1609.92	1609.92	4123.74	**4.91**	413.81

(b) Results for DF–2 ($\approx$ 1000 points)

Table 3

RESULTS FOR DF–1 & DF–2 USING VARIOUS METHODS

DS		StM	ScM	FDM	SM	KM	LHM
DS-1	m_{opt}	**12**	8	8	1	1	1
	E_{NN} (x 10^3)	**4.29**	6.40	6.40	51.00	51.00	51.00
	E_L (x 10^2)	6.65	6.74	6.74	**6.60**	**6.60**	**6.60**
	R (x 10^{-6})	12.62	5.58	5.58	**0.00**	**0.00**	**0.00**
DS-2	m_{opt}	12	8	8	14	4	**33**
	E_{NN} (x 10^3)	4.33	6.48	6.48	3.72	12.94	**1.68**
	E_L (x 10^2)	7.69	10.57	10.57	7.41	10.68	**6.95**
	R (x 10^{-5})	72.26	209.96	209.96	42.42	514.19	**6.35**
DS-3	m_{opt}	12	3	2	17	12	**31**
	E_{NN} (x 10^3)	2.65	10.19	16.79	1.91	2.65	**1.20**
	E_L (x 10^2)	7.64	37.73	144.58	7.14	7.64	**6.99**
	R (x 10^{-4})	12.04	552.52	**0.00**	4.65	12.04	1.77
DS-4	m_{opt}	13	5	3	25	17	**48**
	E_{NN} (x 10^3)	4.74	11.53	17.14	2.54	3.65	**1.40**
	E_L (x 10^2)	10.98	39.56	78.06	7.40	8.57	**7.07**
	R (x 10^{-4})	73.21	410.54	470.35	18.01	39.07	**2.82**
DS-5	m_{opt}	13	8	8	11	7	**17**
	E_{NN} (x 10^3)	4.75	7.64	7.64	5.60	8.66	**3.60**
	E_L (x 10^2)	8.02	10.86	10.86	8.31	11.48	**7.24**
	R (x 10^{-5})	6.38	12.81	12.81	7.21	16.79	**2.34**
DS-6	m_{opt}	13	6	4	23	17	**30**
	E_{NN} (x 10^3)	4.78	10.48	16.10	2.77	3.68	**2.05**
	E_L (x 10^2)	9.44	34.01	72.55	7.24	7.61	**7.01**
	R (x 10^{-5})	46.38	354.75	737.01	12.10	27.92	**3.48**
DS-7	m_{opt}	13	7	6	14	12	**20**
	E_{NN} (x 10^3)	4.78	8.80	10.22	4.41	5.12	**3.11**
	E_L (x 10^2)	8.16	13.66	18.73	7.56	8.34	**7.35**
	R (x 10^{-5})	11.14	99.50	155.56	11.34	15.44	**4.04**
DS-8	m_{opt}	13	7	6	13	13	**19**
	E_{NN} (x 10^3)	4.75	8.69	10.18	4.75	4.75	**3.28**
	E_L (x 10^2)	8.41	16.53	21.76	8.41	8.41	**7.55**
	R (x 10^{-5})	16.78	90.27	141.93	16.78	16.78	**4.30**
DS-9	m_{opt}	12	5	2	38	18	**47**
	E_{NN} (x 10^3)	4.37	9.32	33.70	1.50	2.93	**1.23**
	E_L (x 10^2)	13.95	20.20	236.71	7.05	8.96	**6.96**
	R (x 10^{-4})	190.43	626.38	**0.00**	13.42	82.32	6.93
DS-10	m_{opt}	13	10	11	**45**	23	37
	E_{NN} (x 10^3)	4.88	6.41	5.79	**1.52**	2.78	1.77
	E_L (x 10^2)	14.06	20.45	17.42	**7.00**	7.90	7.06
	R (x 10^{-4})	49.83	73.80	64.99	4.19	14.59	**3.25**
DS-11	m_{opt}	13	7	5	13	13	**34**
	E_{NN} (x 10^3)	4.04	7.55	10.44	4.04	4.04	**1.61**
	E_L (x 10^2)	8.51	14.55	27.88	8.51	8.51	**7.17**
	R (x 10^{-4})	12.27	37.72	35.01	12.27	12.27	**2.17**
DS-12	m_{opt}	13	7	7	8	4	**31**
	E_{NN} (x 10^3)	4.01	7.40	7.40	6.49	12.88	**1.76**
	E_L (x 10^2)	8.35	14.10	14.10	8.38	12.50	**6.93**
	R (x 10^{-6})	1710.36	775.42	775.42	3847.28	**2.82**	208.35

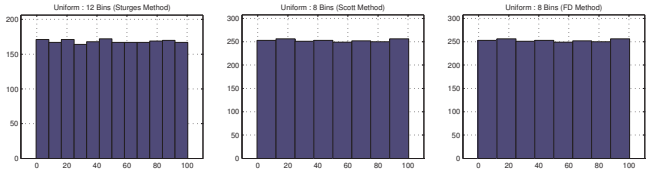
(a) Results for DF–3 ($\approx$ 2000 points)

DS		StM	ScM	FDM	SM	KM	LHM
DS-1	m_{opt}	**14**	6	6	1	1	1
	E_{NN} (x 10^3)	**9.24**	21.29	21.29	127.53	127.53	127.53
	E_L (x 10^3)	1.72	1.72	1.72	**1.72**	**1.72**	**1.72**
	R (x 10^{-6})	9.75	5.31	5.31	**0.00**	**0.00**	**0.00**
DS-2	m_{opt}	14	6	6	18	4	**30**
	E_{NN} (x 10^3)	9.24	21.28	21.28	7.22	31.89	**4.50**
	E_L (x 10^3)	1.84	3.76	3.76	1.73	2.70	**1.69**
	R (x 10^{-5})	39.66	256.35	256.35	18.99	503.03	**5.20**
DS-3	m_{opt}	14	2	2	22	19	**61**
	E_{NN} (x 10^3)	5.68	41.89	41.89	3.74	4.29	**1.35**
	E_L (x 10^3)	1.82	36.47	36.47	1.69	1.70	**1.66**
	R (x 10^{-5})	67.85	**0.00**	**0.00**	15.97	22.60	5.60
DS-4	m_{opt}	14	3	2	54	21	**63**
	E_{NN} (x 10^3)	11.13	42.14	99.51	3.20	7.40	**2.88**
	E_L (x 10^3)	2.84	19.42	90.85	**1.70**	1.90	1.70
	R (x 10^{-4})	49.05	490.78	**0.00**	2.59	25.14	1.84
DS-5	m_{opt}	14	6	6	15	12	**25**
	E_{NN} (x 10^3)	10.82	24.93	24.93	10.12	12.59	**6.17**
	E_L (x 10^3)	1.86	3.31	3.31	1.80	1.91	**1.70**
	R (x 10^{-5})	4.52	23.42	23.42	3.94	4.85	**1.66**
DS-6	m_{opt}	14	4	3	28	17	**40**
	E_{NN} (x 10^3)	11.01	39.73	54.54	5.64	9.03	**3.88**
	E_L (x 10^3)	2.21	18.23	31.41	1.74	1.91	**1.72**
	R (x 10^{-5})	35.02	750.14	1023.42	9.27	20.94	**2.58**
DS-7	m_{opt}	14	5	5	21	16	**28**
	E_{NN} (x 10^3)	10.89	29.98	29.98	7.34	9.54	**5.57**
	E_L (x 10^3)	1.94	6.85	6.85	1.72	1.84	**1.72**
	R (x 10^{-5})	8.72	240.39	240.39	2.63	5.40	**2.43**
DS-8	m_{opt}	14	5	4	15	15	**33**
	E_{NN} (x 10^3)	10.79	29.97	36.79	10.12	10.12	**4.78**
	E_L (x 10^3)	1.95	7.89	10.33	1.90	1.90	**1.72**
	R (x 10^{-5})	11.66	235.58	509.14	10.42	10.42	**2.70**
DS-9	m_{opt}	14	4	2	51	30	**67**
	E_{NN} (x 10^3)	9.12	37.21	83.41	3.05	4.53	**2.48**
	E_L (x 10^3)	3.26	20.41	59.32	1.70	1.80	**1.69**
	R (x 10^{-4})	124.90	79.24	**0.00**	7.60	21.94	2.89
DS-10	m_{opt}	14	7	8	50	30	**59**
	E_{NN} (x 10^3)	11.14	22.96	19.91	3.42	5.32	**2.94**
	E_L (x 10^3)	3.13	9.23	7.15	1.73	1.83	**1.73**
	R (x 10^{-4})	44.75	90.91	92.32	2.16	6.00	**1.74**
DS-11	m_{opt}	14	5	3	26	17	**44**
	E_{NN} (x 10^3)	9.27	25.71	45.70	5.09	7.66	**3.27**
	E_L (x 10^3)	2.27	7.01	14.85	1.76	1.95	**1.71**
	R (x 10^{-5})	84.47	372.90	636.10	31.79	70.25	**8.79**
DS-12	m_{opt}	14	6	5	20	8	**43**
	E_{NN} (x 10^3)	9.20	21.31	25.56	6.54	15.97	**3.35**
	E_L (x 10^3)	2.01	3.99	4.98	1.75	2.08	**1.69**
	R (x 10^{-5})	145.12	30.04	**3.40**	69.38	396.55	8.77

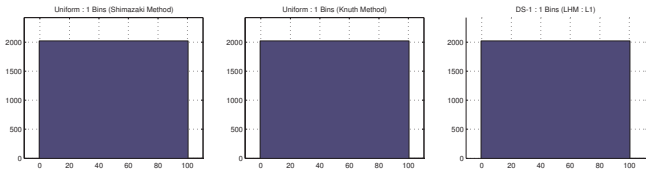
(b) Results for DF–4 ($\approx$ 5000 points)

Table 4

RESULTS FOR DF–3 & DF–4 USING VARIOUS METHODS



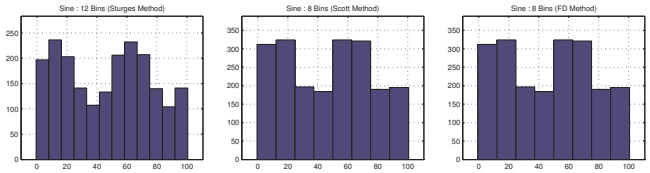
(a) StM (b) ScM (c) FD



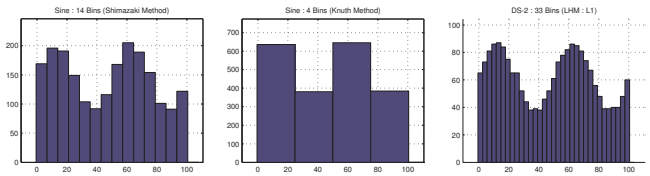
(d) SM (e) KM (f) LHM

Fig. 6

HISTOGRAMS GENERATED FOR DS–1 (FROM DF–3) USING VARIOUS METHODS.



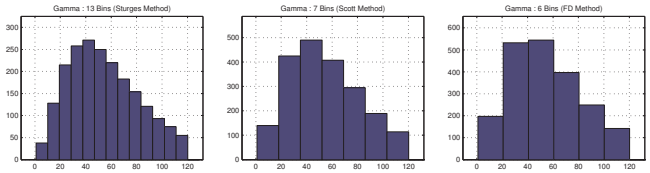
(a) StM (b) ScM (c) FD



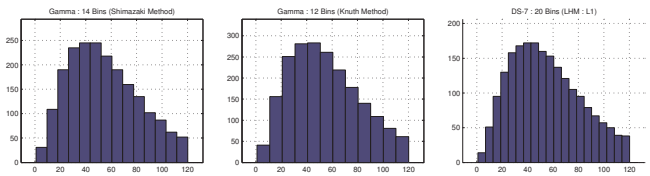
(d) SM (e) KM (f) LHM

Fig. 7

HISTOGRAMS GENERATED FOR DS–2 (FROM DF–3) USING VARIOUS METHODS.



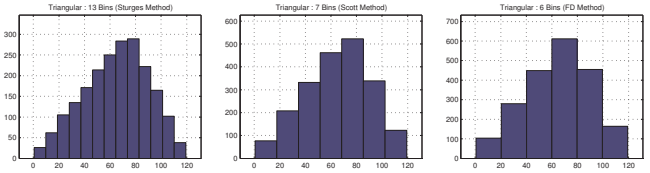
(a) StM (b) ScM (c) FD



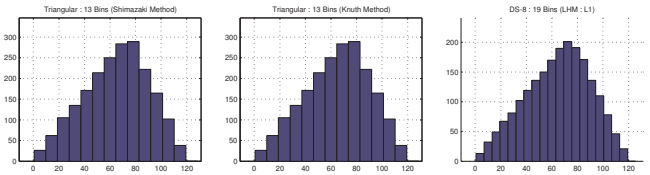
(d) SM (e) KM (f) LHM

Fig. 8

HISTOGRAMS GENERATED FOR DS–7 (FROM DF–3) USING VARIOUS METHODS.



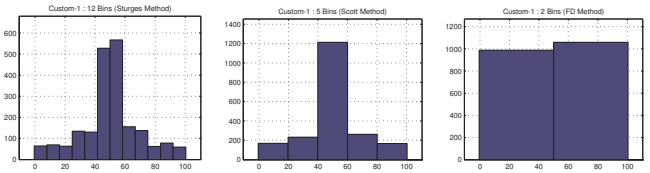
(a) StM (b) ScM (c) FD



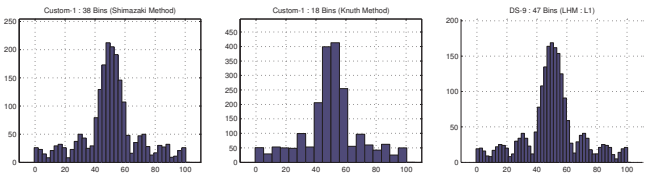
(d) SM (e) KM (f) LHM

Fig. 9

HISTOGRAMS GENERATED FOR DS–8 (FROM DF–3) USING VARIOUS METHODS.



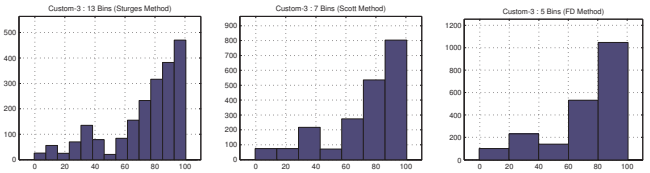
(a) StM (b) ScM (c) FD



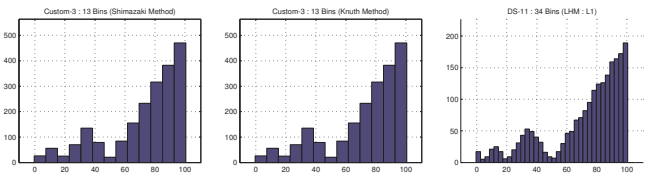
(d) SM (e) KM (f) LHM

Fig. 10

HISTOGRAMS GENERATED FOR DS–9 (FROM DF–3) USING VARIOUS METHODS.



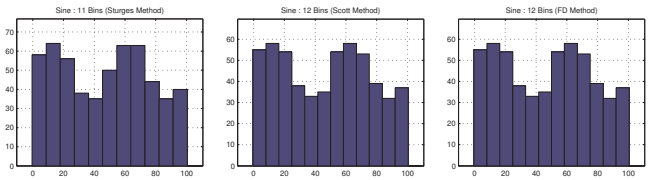
(a) StM (b) ScM (c) FD



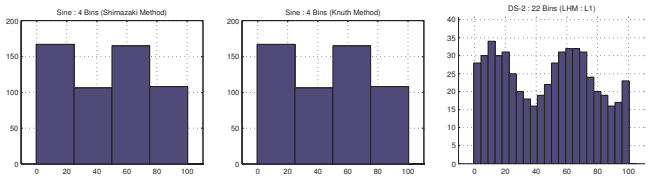
(d) SM (e) KM (f) LHM

Fig. 11

HISTOGRAMS GENERATED FOR DS–11 (FROM DF–3) USING VARIOUS METHODS.



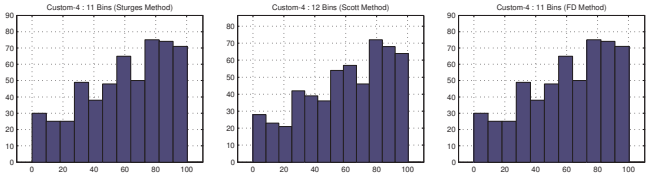
(a) StM (b) ScM (c) FD



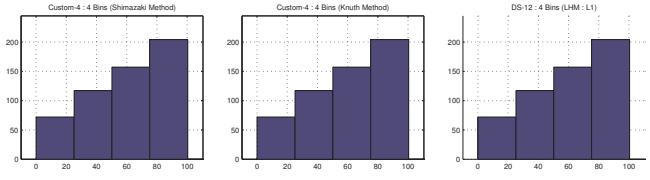
(d) SM (e) KM (f) LHM

Fig. 12

LESS SATISFYING RESULT (LHM): UNDESIRABLE SPIKE ON LEFT MODE (DS–2, DF–1).



(a) StM (b) ScM (c) FD



(d) SM (e) KM (f) LHM

Fig. 13

LESS SATISFYING RESULT (LHM): SHAPE NOT CAPTURED "WELL" (DS–12, DF–1).

5. Conclusions

This paper introduces a new method for selecting the number of bins for constructing a histogram for a given dataset. The performance of the proposed method is compared with the performance of five other methods in the literature. Comparison results show that the proposed method performs better than the other five methods, with the proposed method producing visually appealing histograms that reveal shape features of underlying distribution to a finer detail without admitting excessive noise.

We suggest that future investigations should explore the following issues: (1) Designing a metric to measure the performance of a histogram as evaluated by human perception; (2) Extension of ideas proposed herein to higher dimensional data; and (3) Optimizing the proposed method to reduce time and memory requirements.

References

[1] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, (4):221–424, 1933.

[2] D. Freedman and P. Diaconis. On the histogram as a density estimator:L2 theory. *Probability Theory and Related Fields*, 57(4):453–476, December 1981.

[3] V. I. Glivenko. Sulla determinazione empirica delle leggi di probabilita. *Giornale dell'Istituto Italiano degli Attuari*, (4):92–99, 1933.

[4] P. J. Green and B.W.Silverman. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall/CRC, 1994.

[5] P. Hall. Akaike's information criterion and kullback-leibler loss for histogram density estimation. *Probability Theory and Related Fields*, 85:449–467, 1990.

[6] R. J. Hyndman. The problem with sturges rule for constructing histograms. *Business*, pages 1–2, July 1995.

[7] K. H. Knuth. Optimal Data-Based Binning for Histograms. *ArXiv Physics e-prints*, May 2006.

[8] Lucien Birgé and Yves Rozenholc. How many bins should be put in a regular histogram. *ESAIM: P&S*, 10:24–45, 2006.

[9] J. S. Marron and A. B. Tsybakov. Visual error criteria for qualitative smoothing. *Journal of the American Statistical Association*, 90(430):499–507, 1995.

[10] W. L. Martinez and A. R. Martinez. *Computational Statistics Handbook with MATLAB, Second Edition (Computer Science and Data Analysis)*. Chapman & Hall/CRC, 2 edition, December 2007.

[11] C. R. Rao, E. J. Wegman, and J. L. Solka. *Handbook of Statistics, Volume 24: Data Mining and Data Visualization (Handbook of Statistics)*. North-Holland Publishing Co., 2005.

[12] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982.

[13] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on*, pages 576 – 584, nov. 2004.

[14] D. W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979.

[15] H. Shimazaki and S. Shinomoto. A method for selecting the bin size of a time histogram. *Neural Comput.*, 19(6):1503–1527, 2007.

[16] J. S. Simonoff and F. Udina. Measuring the stability of histogram appearance when the anchor position is changed. *Comput. Stat. Data Anal.*, 23(3):335–353, 1997.

[17] C. J. Stone. An asymptotically optimal histogram selection rule. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983)*, Wadsworth Statist./Probab. Ser., pages 513–520. Wadsworth, 1985.

[18] H. A. Sturges. The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66, 1926.

[19] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

[20] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51:59–64, 1996.

[21] M. P. Wand and M. C. Jones. *Kernel Smoothing (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*. Chapman and Hall/CRC, 1994.

[22] Q. Zhao, M. Xu, and P. Fränti. Knee point detection on bayesian information criterion. In *ICTAI '08: Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence*, pages 431–438, Washington, DC, USA, 2008. IEEE Computer Society.