Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation.

Antonio Irpino*, Elvira Romano**

* Dipartimento di studi europei e mediterranei
Seconda Università degli Studi di Napoli
Via del Setificio, 15 Complesso Monumentale Belvedere - San Leucio
I-81020 Caserta (CE)
irpino@unina.it
** Dipartimento di Matematica e Statistica
Universita' degli Studi di Napoli "Federico II"
Via Cintia - Complesso Monte Sant'Angelo
I-80126 Napoli
elvrom@unina.it

Summary. Histogram representation of a large set of data is a good way to summarize and visualize data and is frequently performed in order to optimize query estimation in DBMS. In this paper, we show the performance and the properties of two strategies for an optimal construction of histograms on a single real valued descriptor on the base of a prior choice of the number of buckets. The first one is based on the Fisher algorithm, while the second one is based on a geometrical procedure for the interpolation of the empirical distribution function by a piecewise linear function. The goodness of fit is computed using the Wasserstein metric between distributions. We compare the proposed method performances against some existing ones on artificial and real datasets.

1 Introduction

Today's storage information mechanism fails to capture a large amount of data and preprocess them in their entirety, while only a summary is stored. In this context *histogram* plays the role of a tool for producing a suitable summarizing description and quickly answering to decision support queries. Following the guide phrase *"An image says more than one hundred words"*, the histogram represents a simple and intuitive graphical tool to describe data distribution. It smoothes the data to display the general shape of an empirical distribution. The problem is that it can give a false impression of the shape of the dataset distribution, because its construction depends on the choice of the number and the length of the subintervals - usually called *buckets or bins* - of the real lines on which the histogram is based. Ideally it could have the situation in which for large bins the nature of the dataset is bimodal and for small bins the plot reduces to unimodal representation. The matter at stake here concerns the *kind of bin width that can take into account the best graphical representation of the underlying DBMS and how it can be constructed with minimal error approximation.*

In database community, and in particular in the framework of query optimization, the search of a good histogram for the representation of a large set of data is best known as the *"selectivity estimation"* problem. Estimates can be used to select the best plan among many competing ones.
There are two main classes of methods for selectivity estimation: sampling methods and statistical methods; in this paper the second kind (nonparametric statistical methods) is taken into account.
In Sec. 2 some of the histogram methods are briefly reviewed, while an excellent taxonomy of histograms can be found in Poosala et al. (1996).
Many data sets have continuous valued attributes such as scientific and statistical data sets. The state-of-the-art histograms implicitly deals with discrete or categorical attribute value domains in which there are relatively few distinct values in the attribute, such methods are used for estimating join selectivities too (see Ioannidis and Poosala (1995)). In the absence of numerous duplicate values in many scientific and statistical data sets, an equi-join will effectively result in the empty set, causing these methods to be ineffective.
Starting from this point, our approach tries to capture statistical variable characteristics, so we can consider it a statistical model based approach, since we aim to approximate (based) the cumulative function by piecewise polynomial (geometrical model).
The proposed methods try to solve the histogram computation in presence of almost continuous datasets according to two different approaches: the first is based on the Fisher algorithm for the partition of ordered data, the latter is based on the best interpolation of the distribution function of data.
The sensitivity of the alternative proposed algorithm is investigated using several dataset and the quality of approximation is computed proposing a goodness of fit measure based on the L_2 Wasserstein metric between two distribution functions. An application on an artificial and on two real dataset is performed in order to corroborate our procedure.

2 Keys proprieties of histogram and a little review of the existing techniques

Let us examine the definition of histogram.

Definition 1
A histogram on a variable X is constructed by partitioning the data distribution into subsets called buckets and approximating the frequencies f and values in each bucket in some common fashion (Ioannidis, 1993).

In this definition it is not mentioned how to draw specific histogram classes and which are the main aspects to consider for its construction. There are six main aspects to be considered in histogram construction: ***Partition Rule, Construction Algorithm, Frequency Approximation, Value approximation, Error Guarantes*** (Ioannidis, 1995).

In the earliest proposed approaches the bin widths were equally spaced and the proposals have been essentially based on choosing the number of bins (Wand, 1997). Nevertheless these methods have the disadvantage of losing the details of high density partition of data. In the last years several types of histogram have been proposed to overcome this problem. In all of them the common guideline is to find the best location of cut points in addition to the number of bind

width to estimate density function (Kooi, 1980). This problem has received attention not only in statistics and database community but also in numerical analysis, where the density function is approximated by a class of polynomial piecewise of some fixed degree.

The common schemes for constructing histograms in DBMS differ in terms of their accuracy of partition constraints. The first proposal goes back to Kooi's PhD Thesis. He introduces a common concept used in statistical literature, the simplest form of histogram, where the value set is divided into ranges of equal length, the so called equi-width histogram. In particular values and frequencies within each bucket, are approximated by the height of the bucket. However, equi-width histograms had not a good improvement over the uniform distribution assumption for the entire value set, that's why new proposals has been done. The so called equi-height or equi-depth histograms (Piatetski-Shapiro, 1984) is one of these. In particular it consists of dividing the set of attribute in buckets that have as approximately the same number of tuples. After these proposals the attention has been shifted to the study of the way in which initial approximation errors is maintained in estimating database through these techniques. The V-optimal histograms (Ioannidis, 1995) have been proposed to minimize the average square error for selectivity estimation problem. In this technique, the partition of the data distribution is computed so that the variance of a source-parameter values within each bucket is minimized. In addition to V-optimal partition constraints, others methods, as this last one, have been developed aiming mainly to group fast several source-parameter values together in the same bucket. Among them, we can distinguish Maxdiff (Ioannidis Y., 1993), which places bucket boundaries between adjacent source-parameter. In addition to these solutions for partition constraints, numerical solution for capturing the shape distribution has received only few attention. Among them it has been proposed to find linear splines for each bin by a least-square regression problem (Konig, 1999), however not much attention has been devoted to the number of parameters to estimate and to the efficient construction cost.

On the basis of the *partition rule*, histograms can be classified according to their mutually orthogonally proprieties (Poosala et al., 1997). The table 1 summarizes and at the same time describes how the existent methods can be collocated. In this frame our method takes place in the context of methods that use the values of observed variable and the relative cumulative frequency.

SORT PARAMETER	SOURCE PAREMETER				
	Spread (S)	Frequency (F)	Area(A)	Cumul. Freq. (C)	Values(V)
Values (V)	Equi-Sum	Equi-Sum V-Optimal Max-.Diff Compressed	V-Optimal Max-.Diff Compressed	Spline-Based V-Optimal ***Piecewise***	***Fisher***
Frequency (F)		V-OPtimal Maxdiff			
Area (A)			Maxdiff		

TAB. 1. – *Map of the main approaches to histogram construction. The algorithms proposed in the present paper are underlined.*

3 The proposed techiques

Let X be a numerical variable whose domain consists of V ordered values.

Let (x_1, x_2,…, x_N) be a list of N observation (tuples) for the variable X, while (v_1, v_2,…, v_V) is the set of distinct values assumed by X in the dataset. The empirical mass function of the X is defined by:

$$f_i = \#\left\{ j : 1 \le j \le N, x_j = v_i \right\} / N$$

We describe the empirical distribution function of X as:

$$c_i = \sum_{j=1}^{i} f_j$$

In the same way we define the mass of an interval $]a,b] \subseteq (-\infty, +\infty)$ as:

$$f(]a,b]) = \#\left\{ j : 1 \le j \le N, a < x_i \le b \right\} / N$$

If the domain is partitioned into β buckets, assuming uniform distribution into the buckets, we calculate the empirical density for the *j-th* ($1 \le j \le \beta$) bucket as:

$$d\left(\left]\underline{b_j}, \overline{b_j}\right]\right) = f\left(\left]\underline{b_j}, \overline{b_j}\right]\right) \Big/ \left(\overline{b_j} - \underline{b_j}\right).$$

The density can be displayed by a histogram, that is a bar graph in which proportion in list are represented by the areas of various bars.

3.1 The Fisher algorithm

The Fisher algorithm can be considered as a V-Optimal(V,V) algorithm. Indeed, the object function that is minimized is the sum of the within buckets variance. V-Optimal algorithm, being fixed a number of buckets β partitioning the domain of values, optimizes a function of the source parameter according to the following formula:

$$\min \sum_{h=1}^{\beta} w_h VAR(X_h)$$

Where X is the source parameter and *w* is a weight of the source parameter.

If the source parameter is the domain of values and *w*=1 it correspond to the minimization of the variance within buckets and leads to the implementation of the dynamic algorithm of partition due to Fisher (1958) for ordered data.

Let us define the following quantities:

$$VAR(k=1) = VAR([v_1, v_V]) = \sum_{i:v_i \in [v_1, v_V]} f_i \left(v_i - AVG([v_1, v_V])\right)^2 \text{ where } AVG([v_1, v_V]) = \sum_{i:v_i \in [v_1, v_V]} f_i v_i$$

The upper bound v_i of the new *k*-th bucket is chosen according to the following dynamic formulation:

$$\underset{v_i}{Arg\min} \left\{ \sum_{h=1}^{H-1} VAR(h) - VAR\left(j \mid j \in (k-1) \text{ and } v_i \in \left[\underline{b_j}, \overline{b_j}\right] \right) + VAR\left(\left[\underline{b_j}, v_i\right]\right) + VAR\left(\left[v_i, \overline{b_j}\right]\right) \right\}$$

where $1 \le j \le H-1 \le k \le \beta$ and $\left[\underline{b_j}, \overline{b_j}\right]$ is the interval notation of the *j*-th bucket.

The computational cost is in terms of operations in the worst case is equal to $O(V^2\beta)$.

3.2 Piece-wise interpolation of the empirical distribution function

The method starts from the trivial histogram -one bucket histogram or the uniform approximation- and at each step the bound of the new bucket is chosen on the base of that value for which it is observed the maximum L_2 distance between the predicted value and the observed value. The distance can be unweighed (in the standard version of the algorithm), or weighted by the number of observations within the buckets. In the second case, if two values have the same error, it is chosen the value in the most populated bucket, according to the practical motivation that it is better to remove an error from a bucket that approximate a large set of observation than from a smallest one.

We start considering the trivial histogram such that:

$$Triv \sim U(v_0, v_V)$$

where v_0 is an artificial point added to the dataset such that $f\left(\left[v_0, v_1\right]\right) = f_1$

In order to identify the best cut point belonging to the v_is into k buckets we solve the following algorithm at each step in order to find β cutpoints. For the standard algorithm (PWst) and for the weighted version (PWw) we have:

$$\underset{v_i}{Argmax}\left\{\left(v_i - v^*_i\right)^2\right\} \text{ (PWst) or } \underset{v_i}{Argmax}\left\{ f(j)\left(v_i - v^*_i\right)^2\right\} \text{ (PWw)}$$

where *f(j)* is the frequency of the *j-th* bucket which includes v_i and where v^*_i is computed by means of the quantile function:

$$v^*_i = \underline{b_j} + \left[c\left(\overline{b_j}\right) - c_i \right] \frac{\overline{b_j} - \underline{b_j}}{f(j)} \;\leftrightarrow\; v_i \in \left[\underline{b_j}, \overline{b_j}\right] \text{ for } j = 1,...,k-1$$

The computational cost in terms of operations is, in the worst case, equal to O(Vβ).

4 The quality measure representation

In the following paragraph we present a more consistent way to compute the mean error square of the obtained histogram and the data distribution according to a metric between distribution.

We develop a measure of accuracy taking into consideration the sum of square differences between the predicted and the observed value, considering the (hypothesised) continuous nature of the model against the discrete observed values.

When we use a continuous function with an histogram (i.e. a mixture of β uniforms with non overlapping supports) to interpolate a discrete right continuous function we always admit an error estimation. Given a vector $[v_1,..,v_i,\ldots,v_V]$ of values with mass function equal to f_i, the best histogram consists of V buckets, and can be represented by a piecewise linear function, where the general linear piece has bounds $(v_i, F(v_i))$ and $(v_{i+1}, F(v_{i+1}))$. The histogram is the best in the sense of piecewise linear interpolation.

Our proposal is to evaluate the procedure accuracy by means of a distance computation between the obtained model of uniforms mixture and the best histogram.

We propose to use the L_2 Wasserstein distance to do the comparison (Gibbs and su, 2002, Barrio et al., 1999, and Verde and Irpino, 2006). It can be considered as the natural extension of the Euclidean distance from point data to distribution data, and it has interesting decomposition properties.

Given two distribution functions F and G, the L_2 Wassertein distance can be computed according to the following formula:

$$d_W^2 = \int_0^1 \left(F^{-1}(t) - G^{-1}(t) \right)^2 dt$$

where F^{-1} and G^{-1} are the quantile functions of the two distributions. The distance computation is heavy when distribution are continuous, but in Verde and Irpino (2006) is showed its feasibility when dealing with histograms.

In Appendix we show that the proposed distance can be decomposed as the sum of the square difference of the means, the square difference of the standard deviations and a residual part that can be assumed as a shape distance between two distributions. The decomposition is summarized as:

$$d_W^2 = \underbrace{\left(\mu_f - \mu_g \right)^2}_{Location} + \underbrace{\left(\sigma_f - \sigma_g \right)^2}_{Size} + \underbrace{2\sigma_f \sigma_g (1 - Corr_{QQ}(F,G))}_{Shape}$$

We consider as maximum error allowed by interpolating data with histogram the L_2 Wasserstein distance between the trivial histogram (the histogram allowing a single bucket) and the optimal one, we may obtain a relative goodness of fit index as the ratio between the square distance of the obtained model (M*) and the optimal histogram (Opt) and the square distance between the trivial (Tri) model and the optimal. We call this measure of as SGFR (Square of Goodness of Fit Ratio):

$$SGFR(M^*) = \frac{d_W^2(M^*, Opt)}{d_W^2(Tri, Opt)}$$

Using the decomposition of the square distance, we may evaluate the "quality" of goodness of fit considering how much of the distance is influenced by a location, a size or a shape difference.

5 An application on real and artificial dataset

We test the proposed techniques on three dataset. The first one is an artificial dataset that derives from the random generation of 10.000 values. It derives from a mixture of three normal distribution *f(x)=0.33*N(*20,20*)+*0.33*N(*40,10*)+*0.34*N(*70,25*).

The second set consists of the 10.000 observations of the variable `dst_bytes` from the `KDD Cup 99` database[1]. This dataset has been chosen for its characteristic of being an example of a peaky (discontinuous) distribution.

The third set collects the first 10.000 observation of the variable `Elevation` from the `Forest cover type` database[2]. This dataset has been chosen for being an example of a smooth (continuous) distribution.

[1] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html

[2] http://kdd.ics.uci.edu/databases/covertype/covertype.html

Artificial	Dataset	Values=10.000	Obs.=10.000			
			Buckets			
Algorithm	Measures	10	25	50	100	200
	Time in sec.	0.55955	0.604214	0.566546	0.58457	**0.56958**
MD	SSE	65.58197	69.49141	75.58185	82.9155	89.14396
	SGFR	1.033287	1.06364	1.109271	1.1618	1.20468
	Time in sec.	40.14913	52.06039	65.02176	82.95111	111.3314
Vopt	SSE	7.71032	2.902464	0.339665	0.13697	0.02490
	SGFR	0.354292	0.217371	0.074347	0.04719	0.02008
	Time in sec.	19.07574	33.5456	55.99695	97.6234	173.8116
FISHER	SSE	8.211255	5.167489	0.955339	0.00252	0.00097
	SGFR	0.36562	0.290044	0.124707	0.00633	0.00386
	Time in sec.	0.240339	0.23878	0.315012	0.52743	1.11027
PWst	SSE	0.512781	0.048029	0.004592	0.00196	0.00052
	SGFR	0.091357	0.027947	0.00862	0.00563	0.00289
	Time in sec.	**0.19953**	**0.23803**	**0.31237**	**0.52245**	1.500763
PWw	SSE	**0.51278**	**0.02326**	**0.00285**	**0.00100**	**0.00036**
	SGFR	**0.09135**	**0.01942**	**0.00676**	**0.00396**	**0.00234**

TAB. 2 – *Synoptics of the performances of the five algorithmes using the first dataset: MD(Maxdiff(F,F)), Vopt (V-Optimal(V,F), Fisher, PWst (Piece Wise approximation to the distribution function, unweighted), PWw (Piece Wise approximation to the distribution function, weighted by the bucket frequency). In bold the best results are showed.*

The comparison methods used here are the MaxDiff(F,F), the V-Optimal(V,F) the Fisher algorithm, the standard and the weighted version of the Piecewise cumulative interpolation algorithm. We measured the operational time using MATLAB^{TM} on a PC (CPU Intel Centrino 1,77Mhz, RAM 1024 MB). The accuracy is computed using the classic error function and the new accuracy measure based on the L_2 Wasserstein metric between distributions.

The main results are collected in tables 2 to 5. While the MaxDiff(F,F) algorithm have the best performances in terms of time spent for the histogram estimation it is less accurate when the number of values of the domain of the variable is very large. The piecewise algorithm is always the best in terms of CPU time and accuracy. To illustrate the enhanced accuracy of the proposed approaches Figure 1 shows the main results for β=10 of the Artificial Dataset.

Concerning the quality of goodness of fit (Tab. 5) the more are the buckets the best all the algorithms (except for the MaxDiff, and the Voptimal(V,F) for the skewed dataset KDD99) fit the first two moments. Fisher algorithm and piecewise ones have better performance over the others. Comparing the quality of goodness of fit between Fisher and piecewise algorithms (Tab. 5), the last seem to be more accurate into the estimation of the first two moments when the number of buckets increases. We suppose that this is due to the fact that the Fisher algorithms, being based on a variance criterion, allow to group data into spherical classes, while piecewise methods are based on the best linear fit of the distribution function, implicitly emphasizing the local uniform density of data.

Forest cov.	Dataset	Values=360	Obs.=10.000			
			Buckets			
Algorithm	Measures	10	25	50	100	200
	Time in sec.	**0.00147**	**0.00144**	**0.00150**	**0.00152**	**0.00165**
MD	SSE	25.28642	19.70946	0.22061	0.07711	0.00861
	SGFR	0.16422	0.14530	0.01547	0.00912	0.00281
	Time in sec.	0.11676	0.24917	0.44977	0.98323	1.37233
Vopt	SSE	0.85188	0.07816	0.07039	0.05140	0.02625
	SGFR	0.03025	0.00849	0.00801	0.00689	0.00481
	Time in sec.	15.11523	30.80159	51.00386	92.46458	165.41046
FISHER	SSE	1.91894	0.22606	0.08548	0.04502	0.02175
	SGFR	0.04506	0.01500	0.00875	0.00603	0.00368
	Time in sec.	0.19329	0.24187	0.32029	0.53857	1.47643
PWst	SSE	**0.44978**	**0.08213**	**0.03813**	0.01217	**0.00153**
	SGFR	**0.02136**	**0.00902**	**0.00596**	0.00323	**0.00105**
	Time in sec.	0.19092	0.24211	0.50761	0.55726	1.20335
PWw	SSE	**0.44978**	0.09306	0.04329	**0.01098**	0.00153
	SGFR	**0.02136**	0.00959	0.00628	**0.00306**	0.00106

TAB. 3 – *Synoptics of the performances of the five algorithmes using the Forest cover type dataset: MD(Maxdiff(F,F)), Vopt (V-Optimal(V,F), Fisher, PWst (Piece Wise approximation to the distribution function, unweighted), PWw (Piece Wise approximation to the distribution function, weighted by the bucket frequency). In bold the best results are showed.*

KDD99	Dataset	Values=2096	Obs.=10.000			
			Buckets			
Algorithm	Measures	10	25	50	100	200
	Time in sec.	0.03753	0.07955	0.02525	0.02647	**0.02837**
MD	SSE	1.2998E+10	1.2516E+09	1.2516E+09	2.6199E+08	1.5893E+08
	SGFR	0.78485	0.24139	0.24139	0.10965	0.08537
	Time in sec.	3.63551	6.94830	11.83113	18.52914	33.68945
Vopt	SSE	8.3791E+09	5.7991E+09	5.7991E+09	1.7037E+09	1.5893E+08
	SGFR	0.62976	0.52340	0.52340	0.28197	0.08537
	Time in sec.	29.57838	73.82461	130.39616	232.45977	309.87441
FISHER	SSE	1.1452E+09	8.0053E+07	4.6098E+04	9.8253E+03	1.8049E+03
	SGFR	0.23301	0.06158	0.00154	0.00068	0.00027
	Time in sec.	0.19037	0.24140	0.34018	0.56922	1.34964
PWst	SSE	2.0296E+06	3.0148E+05	1.9798E+04	4.5786E+03	4.9934E+02
	SGFR	0.00965	0.00370	0.00098	0.00046	0.00015
	Time in sec.	0.19179	0.24546	0.33983	0.56360	1.16070
PWw	SSE	**2.4614E+05**	**3.8040E+04**	**8.3669E+03**	**1.8198E+03**	**221.99776**
	SGFR	**0.00342**	**0.00128**	**0.00052**	**0.00026**	**0.00009**

TAB. 4 – *Synoptics of the performances of the five algorithmes using the KDD 99 dataset: MD(Maxdiff(F,F)), Vopt (V-Optimal(V,F), Fisher, PWst (Piece Wise approximation to the distribution function, unweighted), PWw (Piece Wise approximation to the distribution function, weighted by the bucket frequency). In bold the best results are showed.*

		Artificial		Forest		Kdd 99	
Alg.	Measures	β=10	β=200	β=10	β=200	β=10	β=200
MD	d^2(M,Opt)	65.58186	89.14386	25.23811	0.00737	1.296E+10	1.533E+08
	μ%,	41.5%	71.5%	1.3%	1.6%	47.2%	1.3%
	σ%,	32.5%	10.5%	0.8%	0.2%	47.0%	76.2%
	s%	26.0%	18.0%	97.9%	98.2%	5.7%	22.5%
Vopt	d^2(M,Opt)	7.71016	0.02479	0.85628	0.02167	8.342E+09	1.533E+08
	μ%,	1.1%	0.0%	0.1%	0.2%	31.8%.	1.3%
	σ%,	27.0%	2.0%	1.4%	0.1%	61.5%	76.2%
	s%	71.8%	98.0%	98.5%	99.7%	6.8%	22.5%
FISHER	d^2(M,Opt)	8.21110	0.00092	1.90010	0.01265	1.142E+09	1.565E+03
	μ%,	2.7%	12.0%	0.6%	0.4%	75.1%	5.2%
	σ%,	4.7%	0.8%	0.3%	0.0%	10.6%	0.9%
	s%	92.6%	87.2%	99.1%	99.6%	14.3%	93.8%
T PWst	d^2(M,Opt)	0.51266	0.00051	0.42695	0.00103	1.960E+06	5.026E+02
	μ%,	1.8%	0.3%	14.5%	0.8%	38.7%	3.0%
	σ%,	15.5%	1.2%	6.4%	0.0%	1.1%	0.3%
	s%	82.7%	98.4%	79.1%	99.1%	60.2%	96.7%
PWw	d^2(M,Opt)	0.51266	0.00034	0.42695	0.00105	2.465E+05	1.830E+02
	μ%,	1.8%	0.2%	14.5%	0.8%	5.8%	0.7%
	σ%,	15.5%	1.0%	6.4%	0.2%	10.9%	0.0%
	s%	82.7%	98.8%	79.1%	99.0%	83.3%	99.2%

TAB. 5 – *Synopses: quality of fit goodness between the model and the best histogram, according to the proposed decomposition of the square Wasserstein distance.*

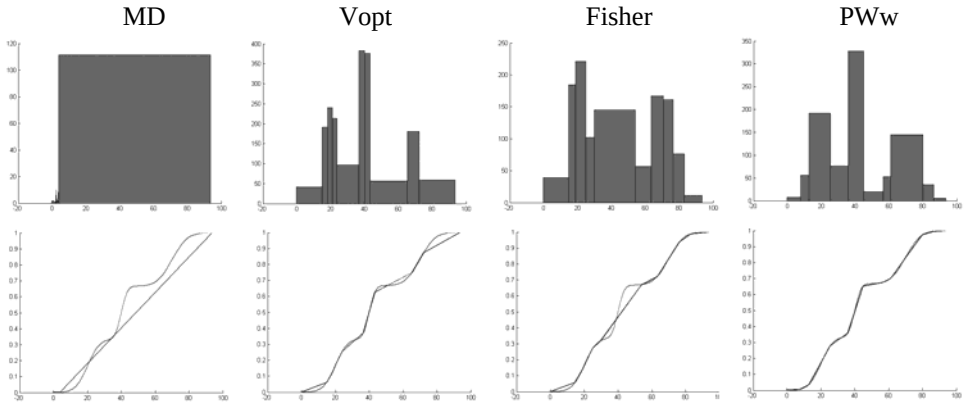

FIG. 1 – *Histogram representation for the Artificial Dataset and illustration of the empirical distribution function approximation for the various methods .*

6 Conclusions and perspectives

In the present paper, several well-established algorithms for the construction of histograms from data have shown to fail in accuracy when the data contained in the database are quasi-continuous, i.e. when the values assumed by the domain of a variable are not so few

with respect the stored tuples. The proposed techniques seems more capable to face this problem. Further, considering the goodness of fit to the best model, the decomposition of the L_2 Wasserstein metric allows to discover the quality of the approximation of a histogram to the data, explaining the distance in terms of goodness of fit of the first two moments and the part of distance due to only a shape factor.

In the present paper the multivariate histogram construction was not considered this will be our next step, naturally in comparison with the existing techniques that seems to suffer the "*curse of dimensionality*" problem.

A deeper insight needs to be given in order to test the proposed techniques in a data stream framework for studying their properties in the case of moving windows histograms or in the case of continuous updates of the histogram models.

References

Barrio, E., Matran, C., Rodriguez-Rodriguez, J. and Cuesta-Albertos, J.A. (1999). Tests of goodness of fit based on the L_2-Wasserstein distance. *Annals of Statistics* (1999), 27, 1230-1239.

Fisher, W.D., (1958). On grouping for maximum homogeneity. *Jorn. Of American Stat. Ass.*, 53, *789-798.*

Gibbs, A.L. and Su, F.E. (2002). On choosing and bounding probability metrics, *International Statistical Review*, 70, 419.

Ioannidis, Y., P. V. (May,1995). Balancing histogram optimality and practicality for query result size estimation. Proc. of ACMSIGMOD, 233-244.

Ioannidis, Y. (1993). Universality of Serial Histograms. *Proceedings of VLDB, Dublin Ireland*, pages 256-277.

Ioannidis, Y. and Poosala, V. (1995). Balancing histogram optimality and practicality for query result size estimation. In *ACM SIGMOD*, 233-244, San Jose, CA.

Konig, A., W. G. (1999). Parametric curve fitting for feedback-driven query result-size estimation. VLDB Conf., 423-434.

Kooi R. (1980). The *Optimization of Queries in Relational Databases*. PhD Thesis, Case Western Reserve University

Piatetski-Shapiro, G., (1984). Accurate estimation of the number of tuples satisfying a condition. Proc. of ACM SIGMOD, 256-276

Poosala V., Ganti V., Ioannidis Y.E. (1999): Approximate Query Answering using Histograms. *IEEE Data Eng. Bull.* 22 (4): 5-14.

Poosala, Y., Ioannidis, Y., Haas, P.J. and Shekita, E.J.(1996). Improved histograms for selectivity estimation of range predicates. In *ACM SIGMOD*, 294-305, Montreal, Canada.

Verde, R., Irpino, A. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data. *Data Science and Classification* (Eds. Batanjeli, Bock, Ferligoj, Ziberna), Springer, Berlin, pp. 185-192.

Wand, M. P. (1997). Data-based choice of histogram bin width.

Appendix

Proof of the decomposition of the Wasserstein distance.

$$d_W^2(F_i(x),F_j(x)) := \int_0^1 \left(F_i^{-1}(t)-F_j^{-1}(t)\right)^2 dt = \\ = \underbrace{\left(\mu_i-\mu_j\right)^2}_{Location} + \underbrace{\left(\sigma_i-\sigma_j\right)^2}_{Size} + \underbrace{2\sigma_i\sigma_j(1-Corr_{QQ}(Y(i),Y(j)))}_{Shape} \quad (1)$$

Let us observe two density functions $f_i(x)$ and $f_j(x)$ having the first two moments finite. To each density function can be associated the distribution functions $F_i(x)$ and $F_j(x)$, the means μ_i and μ_j, the standard deviations σ_i and σ_j where:

$$\mu_i = \int_{-\infty}^{+\infty} x \cdot f_i(x)dx = \int_0^1 F_i^{-1}(t)dt \quad (2)$$

Indeed

$$\int_{-\infty}^{+\infty} xf(x)dx = \int_{-\infty}^{+\infty} xdF(x)$$

if $t=F(x)$ and considering that

$$x = F^{-1}(F(x)) = F^{-1}(t)$$

by substitution we obtain

$$\mu = \int_0^1 F^{-1}(t)dt \quad (3)$$

And where:

$$\sigma^2(x) = \int_{-\infty}^{+\infty} x^2 f(x)dx - (\mu)^2 = \int_0^1 \left(F^{-1}(t)\right)^2 dt - (\mu)^2 \quad (4)$$

for the same substitutions adopted above

Now let assume to centre the two distributions using their means such that:

$$z = x - \mu_i \quad F_i^{-1c}(t) = z \quad F_i^{-1c}(t) = F_i^{-1}(t) - \mu_i \quad (5)$$

In Barrio et al. (1999) is proven that

$$d_W^2\left(F_i(x),F_j(x)\right) := \left(\mu_i-\mu_j\right)^2 + d_W^2\left(F_i^c(x),F_j^c(x)\right) \quad (6)$$

where

$$d_W^2\left(F_i^c(x),F_j^c(x)\right) := \int_0^1 \left(F_i^{-1c}(t)-F_j^{-1c}(t)\right)^2 dt \quad (7)$$

Developing the square we obtain

$$d_W^2\left(F_i^c(x),F_j^c(x)\right) := \int_0^1 \left(F_i^{-1c}(t)\right)^2 dt + \int_0^1 \left(F_j^{-1c}(t)\right)^2 dt - 2\int_0^1 F_i^{-1c}(t)F_j^{-1c}(t)dt =$$

$$= \int_0^1 \left(F_i^{-1}(t)-\mu_i\right)^2 dt + \int_0^1 \left(F_j^{-1}(t)-\mu_j\right)^2 dt - 2\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt = \quad (8)$$

$$= \sigma_i^2 + \sigma_j^2 - 2\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt$$

Let us consider the following quantity

$$\rho_{QQ} = \frac{\int_0^1 F_i^{-1c}(t)F_j^{-1c}(t)dt}{\sqrt{\int_0^1 \left(F_i^{-1c}(t)\right)^2 dt \int_0^1 \left(F_j^{-1c}(t)\right)^2 dt}} = \frac{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt}{\sqrt{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)^2 dt \int_0^1 \left(F_j^{-1}(t)-\mu_j\right)^2 dt}} = \frac{\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt}{\sigma_i\sigma_j} \quad (9)$$

It can be considered as the correlation of two series of data where each couple of observations is represented respectively by the *t-th* quantile of the first distribution and the *t-th* quantile of the second. In this sense we may consider it as the correlation between quantile functions represented by the curve of the infinite quantile points in a QQ plot.

It is worth noting that $0 < \rho_{QQ} \le 1$ differently from the classical range of variation of the Bravais-Pearson's correlation index (-1,+1).

Equation (8) can be rewritten as

$$d_W^2\left(F_i^c(x),F_j^c(x)\right) := \sigma_i^2 + \sigma_j^2 - 2\int_0^1 \left(F_i^{-1}(t)-\mu_i\right)\left(F_j^{-1}(t)-\mu_j\right)dt = \sigma_i^2 + \sigma_j^2 - 2\rho_{QQ}\sigma_i\sigma_j \quad (10)$$

Adding and subtracting $2\sigma_i\sigma_j$ we obtain

$$d_W^2\left(F_i^c(x),F_j^c(x)\right) := \sigma_i^2 + \sigma_j^2 - 2\sigma_i\sigma_j + 2\sigma_i\sigma_j - 2\rho_{QQ}\sigma_i\sigma_j = \left(\sigma_i-\sigma_j\right)^2 + 2\sigma_i\sigma_j\left(1-\rho_{QQ}\right) \quad (11)$$

We may replace this result in (6) obtaining:

$$d_W^2\left(F_i(x),F_j(x)\right) := \left(\mu_i-\mu_j\right)^2 + d_W^2\left(F_i^c(x),F_j^c(x)\right) = \left(\mu_i-\mu_j\right)^2 + \left(\sigma_i-\sigma_j\right)^2 + 2\sigma_i\sigma_j\left(1-\rho_{QQ}\right)$$

QED

Résumé

La représentation d'histogramme d'un grand ensemble de données est une bonne manière pour résumer et visualiser des données et est fréquemment exécutée afin d'optimiser l'évaluation de requêtes dans le système de gestion de bases de données. En cet article, nous montrons les performance et les propriétés de deux stratégies pour une construction optimale des histogrammes sur un descripteur à valeurs réelles sur la base d'un choix apriori du nombre de intervalles élémentaires. Le premier est basé sur l'algorithme de Fisher, alors que le second est basé sur un procédé géométrique pour l'interpolation de la fonction de distribution empirique par une fonction par morceaux linéaire. La qualité de l'ajustement est calculée en utilisant le Wasserstein métrique entre les distributions. Nous comparons les exécutions des méthodes proposées contre quelques celles existantes sur des ensembles de données artificiels et réels.