Rough Histograms for Robust Statistics

Olivier Strauss, Frédéric Comby, Marie-José Aldon
LIRMM - Université Montpellier II - 161 rue ADA, 34392 Montpellier Cedex 5 - FRANCE
Tel. +33 467 418 587 - Fax +33 467 418 500
E- mail: Olivier.Strauss@lirmm.lirmm.fr

Abstract

Applied statistics are widely used in pattern recognition and other computing applications to find the most likely value of a parameter. The use of classical empirical statistics is based upon assumption about normality of underlying density distribution of data. When the data is corrupted by contaminated noise, then classical tools are usually not robust enough and the estimation of the mode is biased. In this article, we propose to estimate the main mode of a distribution by means of a rough histogram and we show that this estimation is robust to contamination.

1. Introduction

Statistics are widely used in different fields of computer science: pattern recognition, filtering, clustering, ... Applied statistics deal with the application of probability theory to provide conclusions that are inferences based on observations. Statistical tools seek to deal with data collection and to reduce this great amount of data into few parameters. Computation of these descriptive parameters is based on the assumption that the probability of an event $\boldsymbol{A}$ can be approximated by the ratio of the number of outcomes that are favorable to $\boldsymbol{A}$ to the number of trials [1].

The problem addressed here is the estimation of the most likely value of a real x knowing n noisy values (x_i) (i=1...n). This most likely value is called the *mode* of the density function of x or mode of x.

If the underlying distribution of x is unimodal, symmetric and not contaminated, then the mode of x corresponds to the mean of x that can be estimated by averaging. Computation of the average is fast and easy. It is widely used in data processing.

When the data are contaminated, then the average is a biased estimate of the mode. Rank statistics (or L-estimates) are a more robust way of performing mode estimation [2]. However, if the underlying distribution is not unimodal, then even L-estimates are biased.

The main mode of the random variable x is the maximum of its density function $f(x)$. This density function can be estimated by constructing a histogram with very small intervals. Then, the main mode is given to be in the interval whose associated accumulator is maximal. However, the size of the interval is limited by the number of data available because of the uncertainty/precision duality. This duality can be set as follows: as the size of the intervals decreases, the precision of the detection of the mode increases while the reliability of the detection also decreases.

In a previous paper [3] we have shown that distributed vote techniques used to build rough histograms are a way of coping with this uncertainty/precision problem. In [4], rough histograms have been used to perform rank statistics (percentile estimate).

In this paper, we propose to use approximate histograms to perform a precise estimation of the main mode of x. This precision can be obtained by disassociating granularity of the histogram and localization of the mode. The granularity of a histogram depends on the histogram quantization, while precision of the mode estimation depends on the number of data and its precision.

The present paper is organized as follows. Section 2 introduces the concept of rough histograms as a generalization of classical (crisp) histograms. Estimation of the main mode of a distribution is presented in Section 3. Some illustrating examples are shown in Section 4. Finally, we provide a short conclusion and discussion on the possible extensions and applications of rough histograms.

2. Rough histograms

2.1. Definitions.

Let (x_i) i=1...N be N real random variables. Computing a histogram of these variables on a real interval $I=[e_{min},e_{max}]$ consists of dividing this interval into p sub-intervals (or cells), and to count the number of x_i that belong to each sub-interval (Fig. 1). The granularity Δ of the histogram, is equal to the cardinality of each sub-interval:

$$\Delta = \frac{e_{max} - e_{min}}{p} \quad (1)$$

An accumulator Acc_k is associated with each cell H_k:

0-7695-0750-6/00 $10.00 © 2000 IEEE

$$Acc_k = \sum_{i=1}^{N} \chi_k(x_i) \quad (2)$$

where $\chi_k(\omega)$ is the characteristic function of the cell H_k.

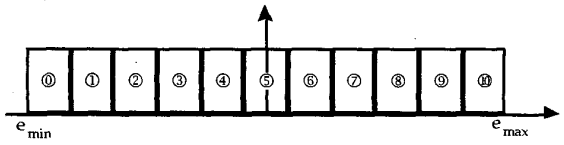


Fig.1: usual partition in 11 cells

To build a rough histogram, the interval I is divided into (p+1) fuzzy subsets (Fig. 2). These subsets make a fuzzy partition of I [5].

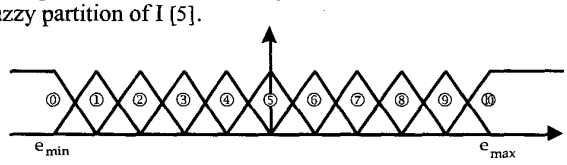


Fig.2: fuzzy partition in 11 cells

Then, the generalization of (2) is:

$$Acc_k = \sum_{i=1}^{N} \mu_k(x_i) \quad (3)$$

This formulation distributes the vote of x_i on two contiguous cells.

When the data is imprecise, this imprecision can be represented by replacing real values x_i by fuzzy or usual intervals X_i. Then (2) must be reformulated as follows:

$$Acc_k = \sum_{i=1}^{N} \underset{\omega \in \Re}{Sup}\left\{\chi_{X_i \cap H_k}(\omega)\right\} \quad (4)$$

and (3) becomes:

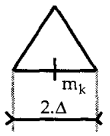
$$Acc_k = \sum_{i=1}^{N} \underset{\omega \in \Re}{Sup}\left\{\mu_{Xi \cap H_k}(\omega)\right\} = \sum_{i=1}^{N} \Pi(X_i; H_k) \quad (5)$$

$\underset{\omega \in \Re}{Sup}\left\{\mu_{X_i \cap H_k}(\omega)\right\}$ is called the possibility of X_i knowing H_k and is written $\Pi(X_i; H_k)$.

2.2. Computation.

When the (x_i) are precise values, then computation of a rough histogram is very simple as shown here. This algorithm uses (p+1) cells H_k (k=0...p).

The core of the cell H_k is $m_k = e_{min} + 2.k.\Delta$ (Δ, given by formula (1), is the spread of the fuzzy set H_k).



The algorithm is:

FOR EACH i
→ IF $x_i < e_{min}$ THEN $Acc_0 = Acc_0 + 1$.
→ ELSE IF $x_n > e_{max}$ $Acc_p = Acc_p + 1$.
→ ELSE COMPUTE $\delta = (x_i - e_{min})/\Delta$, LET $k = trunc(\delta)$ (trunc() is the truncation function) and $\xi = \delta - k$ $Acc_k = Acc_k + (1-\xi)$ AND $Acc_{k+1} = Acc_{k+1} + \xi$.
END FOR.

Computation of rough histograms is not time-consuming and has a computation complexity as low as that of usual histograms.

3. Mode estimation

The granularity of a histogram depends on the x-space quantization. It is defined by Δ, the cardinality of the cells of this histogram ($\Delta = |H_k|$). In crisp histograms, the granularity limits the precision of the mode localization. The estimate of the main mode at granularity Δ is given by the interval H_m such that $\forall k \in [0,p]$, $Acc_k \le Acc_m$. The reliability of this localization is linked to the ratio (Acc_m / N).

The granularity of a rough histogram limits its ability to separate two modes. The precision of the detection only depends on the underlying density distribution.

Finding the main mode of the random distribution (x_i) at granularity Γ consists of finding an interval whose cardinality equals Γ, such that the number Acc(W) of x_i belonging to this interval is maximum compared to any other interval with cardinality Γ.

The idea behind main mode estimation using rough histogram is to estimate Acc(W) for any $W \subseteq I$.

3.1. Linear estimation.

A first approach suggested in [6] consists of using plausibility and credibility measurements of the event $(x \in W)$:

$$Pl(x \in W) = \sum_{k=0}^{p} m(H_k).\Pi(W; H_k) \quad (6)$$

$$Cr(x \in W) = \sum_{k=0}^{p} m(H_k).N(W; H_k) \quad (7)$$

with $m(H_k) = \frac{Acc_k}{N}$.

Plausibility and credibility are respectively upper and lower approximations of $Pr(x \in W)$, the probability that x belongs to W: $Pr(x \in W) \in [Cr(x \in W), Pl(x \in W)]$ (8)

This probability is defined by:

$$Pr(x \in W) = \lim_{N \to \infty} \frac{Acc(W)}{N} \quad (9)$$

Considering (6), (7), (8) and (9), Acc(W) is given by:

$$\sum_{k=0}^{p} Acc_k.N(W; H_k) \le Acc(W) \le \sum_{k=0}^{p} Acc_k.\Pi(W; H_k) \quad (10)$$

Acc(W) can be found to be the solution of a class of relatively simple linear programming problems, by using Dubois-Prade's theorem [6].

Then, finding the maximum consists of building a lower and upper estimation of the density function, then of finding the maximum using statistical reasoning with imprecise probabilities [7].

However, this procedure is time consuming. We are looking for a mode detection method with a complexity as low as that of usual maximum detection methods.

3.2. Pignistic estimation.

A good estimation of the "most likely" value of Acc(W) can be given by the pignistic estimation of Pr(x∈W) [8]:

$$\mathrm{BetP}(W)=\sum_{k=0}^{p} m(H_k).\frac{|W\cap H_k|}{|H_k|} \qquad (11)$$

Then using (9) and (11) Acc(W) can be estimated by:

$$\mathrm{Acc}(W)=N.\mathrm{BetP}(W)=\sum_{k=0}^{p}\mathrm{Acc}_k.\frac{|W\cap H_k|}{|H_k|} \qquad (12)$$

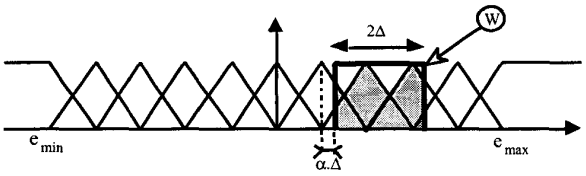


Fig.3: Pignistic transformation

To account for the granularity of the rough histogram, W has to be sought as a subset (fuzzy or crisp) whose cardinality is greater than Δ.
Unfortunately, if W is fuzzy, then the pignistic transformation is not regular, i.e. $\sum_{k=0}^{p}\frac{|W\cap H_k|}{|H_k|}\neq 1$. Thus, this pignistic estimation has to be normalized. Finally, (12) becomes:

$$\mathrm{Acc}(W)=\frac{1}{\lambda}\sum_{k=0}^{p}\mathrm{Acc}_k.\frac{|W\cap H_k|}{|H_k|}=\sum_{k=0}^{p}\mathrm{Acc}_k.\beta_k$$
$$\text{with } \lambda=\sum_{k=0}^{p}\frac{|W\cap H_k|}{|H_k|} \text{ and } \beta_k=\frac{|W\cap H_k|}{\lambda|H_k|} \qquad (13)$$

3.3. Computation of the mode.

In this section, we consider an illustrative example: W is a crisp interval, centered on w, whose spread is 2.Δ: W=[w-Δ, w+Δ]. Then (13) must be rewritten as:

$$\mathrm{Acc}(W)=\sum_{k=0}^{p}\mathrm{Acc}_k.\beta_k=\sum_{k=i-1}^{i+2}\mathrm{Acc}_k\beta_k \qquad (14)$$

where i is the first integer such that $i\le\frac{w-e_{min}}{\Delta}$.

Let $\alpha=\frac{w-e_{min}}{\Delta}-i$ (see fig. 3) then:

$$\beta_{i-1}=\frac{(1-\alpha)^2}{4}\qquad \beta_i=\frac{2-\alpha^2}{4}$$
$$\beta_{i+1}=\frac{1+2\alpha-\alpha^2}{4}\qquad \beta_{i+2}=\frac{\alpha^2}{4} \qquad (15)$$

and $\beta_k=0$ if $k\notin[i-1, i+2]$

Then,

$$\mathrm{Acc}(W)=\frac{1}{4}\left(\begin{array}{l}\alpha^2\left(H_{i-1}-H_i-H_{i+1}+H_{i+2}\right)+\\ 2\alpha.\left(H_{i+1}-H_{i-1}\right)+\left(H_{i-1}+2.H_i+H_{i+1}\right)\end{array}\right) \qquad (16)$$

Formula (16) holds for any set W such that $w\in I_i=\left[e_{min}+i.\Delta, e_{min}+(i+1)\Delta\right]$. Then, if there is a subset $\hat{W}_i$ such that $\mathrm{Acc}(\hat{W}_i)$ is a local maximum, then, derivative of Acc(W) with respect to α must be zero.

$$\frac{\partial(\mathrm{Acc}(W))}{\partial\alpha}=\frac{1}{2}\left(\alpha.\left(H_{i-1}-H_i-H_{i+1}+H_{i+2}\right)+\left(H_{i+1}-H_{i-1}\right)\right) \qquad (17)$$

So $\hat{W}_i$ is a local maximum if

$$\left.\frac{\partial(\mathrm{Acc}(W))}{\partial\alpha}\right|_{\hat{W}_i}=0 \qquad (18)$$

and if $\mathrm{Acc}(\hat{W}_i)>\mathrm{Acc}_i$ and $\mathrm{Acc}(\hat{W}_i)>\mathrm{Acc}_{i+1}$.

$$\hat{W}_i=[e_{min}+(\hat{\alpha}_i-1)\Delta, e_{min}+(\hat{\alpha}_i+1)\Delta] \qquad (19)$$

$$\text{with: } \hat{\alpha}_i=\frac{H_{i-1}-H_{i+1}}{H_{i-1}-H_i-H_{i+1}+H_{i+2}} \qquad (20)$$

Thus, finding the main mode consists of finding $\hat{\alpha}_i$ for each cell of the rough histogram and select the cell H_m with $\mathrm{Acc}(\hat{W}_m)\ge\mathrm{Acc}(\hat{W}_i)$ $(i\neq m)$. $\bar{x}$, the estimate of the mode of x, is then given by:

$$\bar{x}=e_{min}+(m+\hat{\alpha}_m)\Delta \qquad (21)$$

4. Application.

In this section, we consider a simple illustrative example. x is a random real variable obtained by simulation of a non symmetric distribution with mode equal to 3.1. This distribution is 0.3-contaminated by a normal and a uniform distribution (fig. 4).

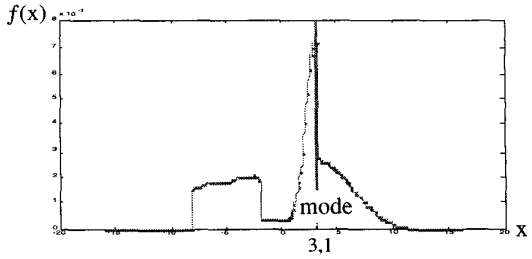


Fig. 4: density distribution of x.

The different parameters are: e_{min}=-15, e_{max}=15, p=20, N=200. The interval I=[-15,15] has been divided into r=1000 samples. Each sample of I is denoted w_i (i=1...r). An interval W_i ($W_i=\left[w_i-\Delta, w_i+\Delta\right]$) is associated with every w_i.

$Acc(W_i)$ has been computed using three different ways:

1) counting the number of x_i that belongs to W_i (empiric density function estimate),

2) computing upper and lower bounds of $Acc(W_i)$ using linear estimation,

3) computing pignistic estimation.

Results of these computations are shown in fig. 5.

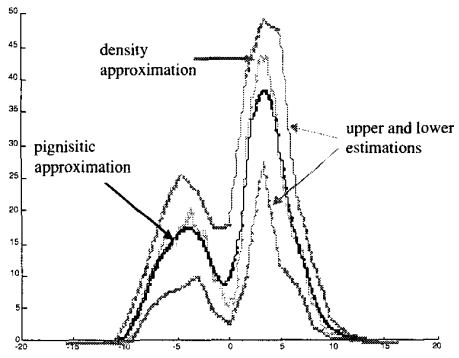


Fig. 5: superimposed density function approximation, upper and lower bounds and pignistic. N=200.

As can be observed, the pignistic transformation looks like a smoothed estimation of empiric density estimate. Both density and pignistic estimations belong to the interval defined by upper and lower linear estimations. At this point, the main difference between pignistic estimation and empiric density estimation is the computation time. Empiric density estimation requires r.N elementary instructions while the pignistic estimation only requires N+r instructions.

Fig. 6 shows the same experiment with N=40. It can be seen that the overall shape of the pignistic estimation doesn't change with N and looks like a quadratic interpolation of the histogram.

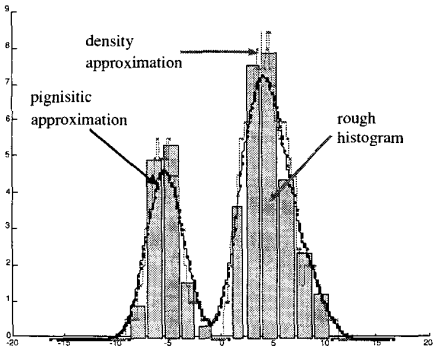


Fig. 6: superimposed density function approximation, pignistic approximation and rough histogram N=40.

Fig. 7 shows the estimation of $\tilde{x}$, mode of x, using averaging, ranking and histograms. Because of contamination, the mean and the median provide biased estimates, while the histogram estimate is more robust.

5. Conclusion and discussion.

In this paper, we have presented a method using rough histograms to estimate the main mode of a distribution. This estimate seems to be more robust than classical methods when the data is corrupted by contaminated noise. Future work will deal with extension of this method to n-dimensional spaces to perform movement or shape detection in video images. In this case, however, a pignistic estimation will limit the method. Computation of 1-dimensional data implies surface calculation, n-dimensional data will involve hyper-volume calculation. In addition, the precision of the data has to be taken into account by associating with each cell of the histogram a complementary accumulator based on a conditional necessity measure.

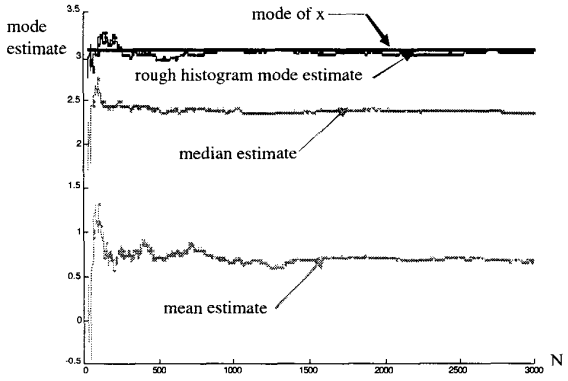


Fig. 7: comparison between mean, median and rough histogram's mode estimates.

6. References

[1] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd edition, McGraw-Hill 1991.

[2] P.J. Hubert, *Robust Statistics*, J. Wiley and sons, 1981.

[3] O. Strauss, "Use the Fuzzy Hough transform: towards reduction of the precision/uncertainty duality", *Pattern Recognition*, vol. 32, n°11, November 1999, pp. 1911-1922,

[4] O. Strauss, "Rough histograms: use for rank statistics", *LFA'99*, Valenciennes October 21,22, 1999.

[5] D.Dubois, H. Prade, "Rough fuzzy sets and fuzzy rough sets", *International Journal of General Systems*, n°17, pp. 191-200, 1990.

[6] D.Dubois, H. Prade, *Possibility theory: an approach to the computerized processing of uncertainty*, Plenum Press, New York, 1981.

[7] P. Walley, *Statistical Reasoning with Imprecise Probabilities*, Chapman and Hall, London, 1991.

[8] T. Denœux, "Modeling vague beliefs using fuzzy-valued belief structures", to appear in *Fuzzy Sets and Systems* (accepted in august 1998).