Information Sciences 298 (2015) 136–158

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

# Probability density function estimation with the frequency polygon transform

Ezequiel López-Rubio *, José Muñoz-Pérez

*Department of Computer Languages and Computer Science, University of Málaga, Bulevar Louis Pasteur, 35, 29071 Málaga, Spain*

ARTICLE INFO

*Article history:*
Received 11 April 2014
Received in revised form 16 October 2014
Accepted 3 December 2014
Available online 12 December 2014

*Keywords:*
Probability density function estimation
Nonparametric estimation
Multivariate frequency polygon
Computer vision
Object tracking

ABSTRACT

Most current nonparametric approaches to probability density function estimation are based on the kernel density estimator, also known as the Parzen window estimator. A usual alternative is the multivariate histogram, which features a low computational complexity. Multivariate frequency polygons have often been neglected, even though they share many of the advantages of the histograms, while they are continuous unlike the histograms. Here we build on our previous work on histograms in order to propose a new probability density estimator which is based on averaging multivariate frequency polygons. The convergence of the estimator is formally proved. Experiments are carried out with synthetic and real machine learning datasets. Finally, image denoising and object tracking applications are also considered.



## 1. Introduction

Probabilistic approaches to pattern recognition and machine learning often involve the estimation of the underlying probability density function (pdf) for a given dataset. Density estimators have been used for many years by statisticians, scientists in general and engineers as tools to draw inferences from data in social, physical and computer sciences [48,56,25]. From the theoretical point of view, density estimation is a complex problem since it is known that all nonparametric estimators of the pdf are biased [49].

As said before, the approximation of the underlying pdf of the given data set arises in many fields. Bayesian decision theory relies on the accurate estimation of the class pdfs [66,8]. Risk assessment and fault detection methods require the estimation of the pdf for historical data in order to predict events [1]. Bayesian image filtering can use pdf estimation to compute the likelihood of a possible original image, given the observed image [14]. Outlier detection and feature selection can be related to the ratio of two pdfs [59,58]. Blind signal detection greatly benefits from the estimation of the pdf of the noise when it is non Gaussian [43]. The estimation of the pdf of pixel values is of paramount importance to some methods for image denoising [46] and medical image segmentation [57]. Finally, the mean shift algorithm finds modes of estimated pdfs. It is in widespread use for object tracking [9,5,13,36,37] and image segmentation [20,24,38,60,67] applications.

Here we develop a continuous pdf estimator for multivariate data with a reduced computational complexity and a high estimation accuracy. It is based on our previous work on multivariate histograms [39], and it addresses the issues which arise when they are replaced by frequency polygons. These issues include the increased computational complexity with respect to

* Corresponding author.
*E-mail addresses:* ezeqlr@lcc.uma.es (E. López-Rubio), munozp@lcc.uma.es (J. Muñoz-Pérez).
*URL:* http://www.lcc.uma.es

http://dx.doi.org/10.1016/j.ins.2014.12.014
0020-0255/

histograms due to the use of more than one bin to estimate the pdf at each test point, the need for a fast procedure to locate these bins, and the fact that the optimal bin widths for histograms and frequency polygons are different. Our proposal yields a smoother and more accurate estimation, and we prove its convergence to the true pdf. Moreover, a mode finding algorithm based on this estimator is proposed as an alternative to the well known mean shift algorithm. This proposal is much faster than the algorithm that was proposed for multivariate histograms and it has less tunable parameters, so it is easier to use in applications.

The structure of this paper is as follows. An overview of related work is done in Section 2. After that. Section 3 provides a gentle introduction to multivariate frequency polygons, so that the context of our proposal is outlined. The proposed pdf estimator and the mode finding algorithm are defined in Section 4. The most important properties of the proposal and the main results of the experiments are discussed in Section 5. Experiments are reported in Section 6. Finally, Section 7 is devoted to conclusions.

## 2. Related work

Parametric estimators based on mixture models can be used to approximate a pdf when the dataset is known to contain some clusters, so that each mixture component models one of these clusters [53]. Here we are interested in the remaining situations, where nonparametric approaches are more suitable. Kernel estimators are by far the most popular [26]. Their fundamental tunable parameter is the kernel bandwidth, whose selection has been extensively studied [35,10,4,2]. They share a computational complexity issue, since in principle all the training samples are retained in the model so as to place a kernel on each sample. This means that the time complexity grows linearly with the number of training samples $N$ and with the number of test samples $M$, so that the estimator is $O(NM)$. Fortunately optimization techniques are available, and among them the Fast Gauss Transform or FGT [22], and its enhancement the Improved Fast Gauss Transform or IFGT [65,42] have become a standard. Furthermore, specialized architectures based on field-programmable gate arrays (FPGAs) have been built to achieve real-time pdf estimation for particularly complex problems [16]. Another strategy to reduce the computational load of the kernel estimator is the use of fast nearest neighbors algorithms based on binary trees, such as kd-trees [6,21], and anchors hierarchies [44]. All of these optimizations produce approximations to the original kernel estimator, so the accuracy of the approximation is an issue to be taken into account.

Multivariate histograms have a fundamental advantage over the kernel estimator, namely its summarizing property. That is, once the training samples have been tallied according to the histogram bins, the complexity of the test phase is no longer dependent on the size of the training set, but only on the number of nonempty bins. This comes at the price that the obtained pdf estimator is not continuous. This makes them unsuitable for many applications, so that their use is focused towards those situations where real time operation is critical. Sparse coding of object appearance is a typical example from computer vision [63,17]. The choice of the bin width plays a role similar to that of the bandwidth of the kernel estimator [31,15].

Frequency polygons are piecewise linear (first order) estimators, so that the discontinuities of the histograms are not present [52,29]. Histograms are piecewise constant (zeroth order) estimators. Frequency polygons have gone almost unnoticed for the practitioners in the above mentioned fields, even if their accuracy is much higher than that of the histograms, while they share their summarizing property [55].

## 3. Motivation

Next a simple example is used in order to highlight the differences between multivariate histograms and frequency polygons. Let us consider the uniform distribution over a circle, from which $N = 1000$ training samples are drawn. For a bin width $h = 0.07$, the training samples are binned into squares of size $0.07 \times 0.07$ (Fig. 1).

At this point two alternatives are possible:

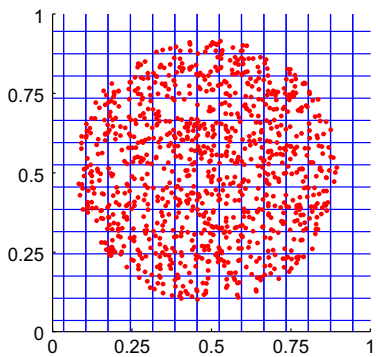


**Fig. 1.** Example training dataset with $N = 1000$ samples plotted as dots. The lines are the bin boundaries for bin width $h = 0.07$.