Assessment of probability density estimation methods: Parzen window and Finite Gaussian Mixtures

C. Archambeau
DICE
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
archambeau@dice.ucl.ac.be

M. Valle
DIBE
Università degli studi di Genova
Genova, Italy
maurizio@micro.dibe.unige.it

A. Assenza
DIPTEM
Università degli studi di Genova
Genova, , Italy
alex.assenza@gmail.com

M. Verleysen
DICE
Université Catholique de Louvain
Louvain-la-Neuve, Belgium
Verleysen@dice.ucl.ac.be

Abstract—**Probability Density Function (PDF) estimation is a very critical task in many applications of data analysis. For example in the Bayesian framework decisions are taken according to Bayes' rule, which directly involves the evaluation of the PDF. Many methods are available to this aim, but there is no consensus in the literature about which to use, nor about the pros and cons of each of them. In this paper we present a thorough and extensive experimental comparison between two of the most popular methods: Parzen window and Finite Gaussian Mixture. Extended experimental results and application development guidelines are reported.**

I. INTRODUCTION

Numerical data are found in many applications of data analysis. A random variable is the mathematical concept that characterizes the numerical results of experiments. To analyse data, one may choose to handle directly the results of the experiments. For example, simple data analysis methods like the linear PCA (Principal Component Analysis), the non-linear MLP (Multi-Layer Perceptron), and many others, work directly on the numerical values of samples. While this way of working may reveal adequate in many situations, other require working with the underlying random variable instead of the numerical sample. A random variable is completely characterized by its Probability Density Function (PDF), i.e. a function whose value in x gives the probability of an event to occur when the random variable is equal to x. Estimating PDF based on samples is of primary importance in many contexts. There exists a lot of methods aimed to estimate PDFs; while all of them are applicable in the univariate case, some of them may also be applied to the multivariate one. However, there is no consensus in the literature about which method to use, nor about the pros and cons of these methods. The aim of this paper is not to answer all the questions regarding the use of the methods for estimating PDF. Nevertheless, it aims at giving some insights into the most effective methods that are traditionally used by data analysts. Popular methods used to estimate PDFs in the experiments are Parzen windowing and Finite Gaussian Mixtures (FGM) that will be described in Paragraph II. Paragraph III will present the adopted experimental procedure. In Paragraph IV the results of some selected experiments are illustrated. Finally, Paragraph V introduces some guidelines for the use of PDF estimation methods.

II. PDF ESTIMATION

Probability density function estimation is a fundamental step in statistics as it characterizes completely the "behaviour" of a random variable. It provides a natural way to investigate the properties of a given data set, i.e. a realization of the random variable, and to carry out efficient data mining. When we perform density estimation three alternatives can be considered. The first approach, known as parametric density estimation, assumes the data is drawn from a specific density model. Unfortunately, an a-priori choice of the PDF model is in practice not suited since it might provide a false representation of the true PDF. An alternative is to build nonparametric PDF estimators [1][2] in order to "let the data speak for themselves". A third approach consists in using semi-parametric models [3]. As nonparametric techniques, they do not assume the a priori shape of the PDF to estimate. However, unlike the nonparametric methods, the complexity of the model is fixed in advance, in order to avoid a prohibitive increase of the number of parameters with the size of the data set. Finite mixture models are commonly used to serve this purpose. In this section, we briefly recall the Parzen window estimator (one of the most representative nonparametric PDF estimators), and show how the kernel width can be selected a priori in the case of Parzen. Next, we present Finite Gaussian mixture models. We refer to X as a continuous random variable and to $p(x)$ as its PDF. Let consider $\{x_n\}_{n=1}^{N}$, a realization of X.

0-7803-9390-2/06/$20.00 ©2006 IEEE

A. Parzen Window

The Parzen window estimator [4] does not assume any functional form of the unknown PDF, as it allows its shape to be entirely determined by the data without having to choose a location of the centers. The PDF is estimated by placing a well-defined kernel function on each data point and then determining a common width c, also denoted as the smoothing parameter. In practice, Gaussian kernels are often used:

$$N\left(x\middle|c,\sigma\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-c)^2}{2\sigma^2}\right)$$

where c and c are the kernel centre and the kernel width respectively. The estimated PDF is then defined as the sum of all the Gaussian kernels, multiplied by a scaling factor:

$$\hat{p}(x)=\frac{1}{N}\sum_{n=1}^{N}N\left(x\middle|x_n,\sigma\right)$$

1) Silverman's plug-in principle

Choosing the kernel width a priori is definitely not the best choice for c, as its optimal value (i.e. the value that minimises the measure of dissimilarity between the estimated PDF and the estimating one) strongly depends on the type of data we are dealing with, their number and the amount of noise they are corrupted by. Silverman proposed in [2] to plug a Gaussian distribution to approximate $p(x)$, leading to the following rule of thumb:

$$\sigma_{\mathrm{SIL}}=0.9AN^{-\frac{1}{5}} \quad \text{with} \quad A=\min\left\{s,\frac{R}{1.34}\right\}$$

In this equation s is the empirical standard deviation and R is the sample interquartile range.

2) Leave-one-out estimator

In order to estimate the optimal value of the kernel width, the leave-one-out approach can be used. It is based on the minimisation of a dissimilarity measure, namely the *integrated square error* (ISE) defined as follows:

$$\mathrm{ISE}=\int\left\{\hat{p}(x)-p(x)\right\}^2dx$$

This error criterion can also be rewritten as follows:

$$\mathrm{ISE}=\int\hat{p}(x)^2dx-2\mathrm{E}\{\hat{p}(x)\}+\int p(x)^2dx$$

Observing that the last term does not depend on c, it can be ignored as far as minimization is concerned. This leaves us with only the second term depending on both c and the unknown density $p(x)$. Seeing that the second term can be approximated by its leave-one-out estimator $\hat{p}_{-n}(x_n)$ [1], we may define the following error criterion:

$$\begin{aligned}\mathrm{E}_{\mathrm{LOO}}(\sigma) &= \int\hat{p}(x)^2dx-2\mathrm{E}\{\hat{p}_{-n}(x_n)\}\\ &\approx \int\hat{p}(x)^2dx-\frac{2}{N}\sum_{\mathrm{n}=1}^{\mathrm{N}}\hat{p}_{-n}(x_n)\end{aligned}$$

where $$\hat{p}_{-n}(x)=\frac{1}{N-1}\sum_{\mathrm{m}=1,\mathrm{m}\neq\mathrm{n}}^{\mathrm{N}}N\left(x\middle|x_m,\sigma\right)$$

Substituting the Parzen estimate of $p(x)$ in this expression, we can obtain the leave-one-out cross-validation criterion:

$$E_{LOO}(\sigma)=\frac{1}{N^2}\sum_{n=1}^{N}\sum_{m=1}^{N}N\left(x_m\middle|x_n,\sqrt{2}\sigma\right)-\frac{2}{N(N-1)}\sum_{n=1}^{N}\sum_{m=1,m\neq n}^{N}N\left(x_n\middle|x_m,\sigma\right)$$

Interestingly, this criterion does not require the evaluation of an integral anymore. Finally, by scanning a certain range of σ the optimal width can be selected: $\sigma_{\mathrm{LOO}}=\arg\min_{\sigma}E_{LOO}(\sigma)$

B. Finite Gaussian Mixtures

Finite mixture distributions can approximate any continuous PDF, provided the model has a sufficient number of components and provided the parameters of the model are chosen correctly [5]. The true PDF is approximated by a linear combination of K component densities:

$$\hat{p}(x)=\sum_{k=1}^{K}P(k)p(x|k)\text{, with } K<<N$$

In this equation $p(x|k)$ is the probability of x given the component distribution k and $P(k)$ are the mixture proportions or priors. The priors are non-negative and must sum to one. In practice, Gaussian kernels are often used: $p(x|k)=N\left(x\middle|c_k,\sigma_k\right)$. A popular technique for approximating iteratively the maximum likelihood estimates of the model parameters $P(k)$, c_k and σ_k is the *expectation-maximization* (EM) algorithm [6]. Let us define the likelihood function:

$$\mathrm{L} = \prod_{n=1}^{N} \hat{p}(x_n)$$

Maximizing the likelihood function is equivalent to finding the most probable PDF estimate provided the data set $\{x_n\}_{n=1}^{N}$. The EM operates in two stages. First, in the *E-step*, the expected value of some "unobserved" data is computed, using the current parameter estimates and the observed data. Here the "unobserved" data are the data labels of the samples. They correspond to the identification number of the different mixture components and specify which one generated each datum. Subsequently, during the *M-step*, the expected values computed in the *E-step* are used to update the model parameters accordingly. Each iteration step can be summarized as follows [3]:

a) *E-step*, $P^{(i)}(k|x_n) = \dfrac{p^{(i)}(x_n|k)P^{(i)}(k)}{\hat{p}^{(i)}(x_n)}$

b) *M-step*, $\mathbf{c}_k{}^{(i+1)} = \dfrac{\Sigma_{n=1}^{N} P^{(i)}(k|x_n)x_n}{\Sigma_{n=1}^{N} P^{(i)}(k|x_n)}$

c) $(\sigma_k^2)^{(i+1)} = \dfrac{\Sigma_{n=1}^{N} P^{(i)}(k|x_n)\left(x_n - c_k{}^{(t+1)}\right)^2}{\Sigma_{n=1}^{N} P^{(i)}(k|x_n)}$

d) $P^{(i+1)}(k) = \dfrac{1}{N} \cdot \sum_{n=1}^{N} P^{(i)}(k|x_n)$

Note that in the value $P^{(i)}(k|x_n)$ computed in the *E-step* corresponds to the posterior probability that a known data sample x_n was generated by component *k*.

III. METHODOLOGY FOR ESTIMATION METHODS COMPARISON

The comparison of the two proposed methods for PDF estimation requires taking some decisions according with the test case and dissimilarity measures to be used and to the operative experimental procedure. We summarise them as follows:

- Identify a reference PDF: the reference PDF represents the test case for estimation methods comparison. It should be affected by the most common singularities and characterised by some of the main issues that should be faced during real world PDFs estimation.

- Set the dimension of the synthetic measures set: the cardinality of the set of measures is a key performance factor: it strongly affects the performance of the PDF estimation methods and it is to be related in some way to the information that the researcher provides the method with. It is a simulation parameter to be set.

- Select a criterion to measure the distance between the reference PDF and the estimating one: one measure of dissimilarity should be selected in order to assess the performance of the estimation methods under comparison.

- State a clear experimental procedure and perform each task: according with the previously defined criterion, experiments should be set in order to perform the comparison with the optimal simulation parameter settings for each single method. Moreover, in order to guarantee a certain degree of generalisation exogenous factors should be taken into account during the statement of the experimental procedure. It is the case of the cardinality of the set of synthetic measures.

A. Rationale of the selected PDF

According with the generalization requirements we aim to, a PDF has been selected by considering some of the most common issues to be faced during real world PDF estimation. It is the case of a PDF resulting from the mixture of multiple parametric PDFs, with two peaks (a higher one and a lower one) forced in the structure and shape of the selected PDF. With more details, the adopted PDF is a mixture of four Gaussian PDFs (i.e. $N_1(x|3.5,1.6)$, $N_2(x|7.5,2)$, $N_3(x|12,2)$, $N_4(x|16.5,1.5)$, where $N_i(x|\mu,\sigma)$ refers to the i^{th} Gaussian density function with mean and standard deviation respectively equal to μ and σ) and one Gamma PDF with α parameter and β respectively equal to 8 and 0.2.

B. Distance measure

To asses the performance we used the Hellinger distance [7]. If we set: a) $S \subseteq \Re$, $a,b \in S$; b) $x \in S$ is a continuous random variable; c) *P(x)* is the probability law of *x*, then the Hellinger distance can be defined by the following expression:

$$H^2(P_1(x), P_2(x)) = \frac{1}{2}\int_a^b \left(\sqrt{P_2(x)} - \sqrt{P_1(x)}\right)^2 dx$$

where the notation $H^2(P_1(x), P_2(x))$ means that this is the formula of the squared Hellinger distance.

IV. EXPERIMENTAL RESULTS

In Parzen's window we can set only the sigma parameter and there isn't standard deviation because is a deterministic method. In the experiments we consider from 0.15 to 1.95 step 0.1 and 2 as value for sigma. Figure 1 shows that σ in Parzen's estimator should be naturally large for small N and small for large N. We made experiments also considering the Silverman's plug-in principle and we obtained good results only with few data according to literature [1][2]. Therefore from experimental results we know that Silverman's plug-in should be a good way to set the sigma parameter when less than 250 data are considered. In FGM we can set the number of kernels and the number of iterations. In the first part of the study we made experiments with 15000 data and different number of iterations and kernels; experiments show (Figure 2) that best results with 200 iterations are obtained independently from the number of kernels. Indeed from this first part of the experiments, we deduced to use in the following experiments always 200 iterations because they give best results without a very large computational cost. In the second part of the experiments with different number of data, we met the problem that EM algorithm collapses with few data and a big number of kernels. Results are shown in Figure 3. FGM requires an initialization that in our experiments was made through the vector quantization [8] with the k-means algorithm. Indeed for FGM we obtain a standard deviation and results show that it's quite small in all configurations.

V. CONCLUSIONS

The goal of our work is to experimentally assess the performance of two popular methods for PDF estimation: Parzen window and FGM. We implemented an exhaustive experimental campaign whose results are summarized in Figure 4. In conclusion if a low number of data available is not encountered, FGM are thus expected to perform well; in particular FGM perform better than the other methods when there are more than 80 data and using 4 kernels. When the number of data is really big, i.e. more than 7500 data, it's better to increase the number of kernels at 15 to obtain the minimum error. Instead with few data, less than 80, FGM don't work but Parzen with Silverman's plug-in is a good estimator; moreover it's easy to use for setting the parameters and the low computational cost. Future work will concern the extension of the comparison to other methods as Vector Quantization.

ACKNOWLEDGMENT

The authors thank Filippo Queirolo for his help in the early stages of this project.

REFERENCES

[1] W. Härdle, M. Müller, S. Sperlich, A. Werwatz: Nonparametric and Semiparametric Models. Springer, New York (2004).

[2] B. W. Silverman: Density Estimation. Chapman & Hall/CRC, London (1986).

[3] G. McLachlan and D. Peel: Finite Mixture Models. Wiley, New York (2000).

[4] E. Parzen, On Estimation of a Probability Density Function and Mode, Annals of Math. Statistics, 33: 1065–1076, 1962.

[5] C. M. Bishop: Neural Networks for Pattern Recognition. Oxford University Press, Oxford (1995).

[6] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM algorithm, J. Roy. Stat. Soc. (B), 39: 1–38, 1977.

[7] Michèle Basseville: Rapports de recherché N°899: Distance measures for signal processing and pattern recognition, INRIA-RENNES, September 1988

[8] S. C. Ahalt, A. K. Krishnamurthy, P. K. Chen, D. E. Melton, Competitive Learning Algorithms for Vector Quantization, Neural Networks, 3(3): 277–290, 1990.

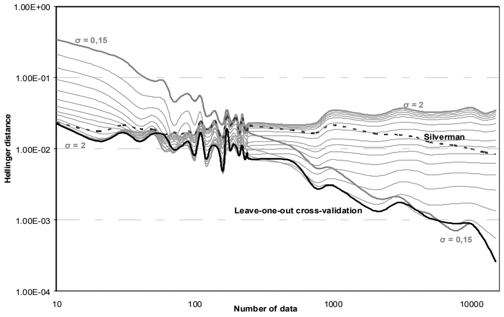

Figure 1. Distance vs number of data, σ as parameter

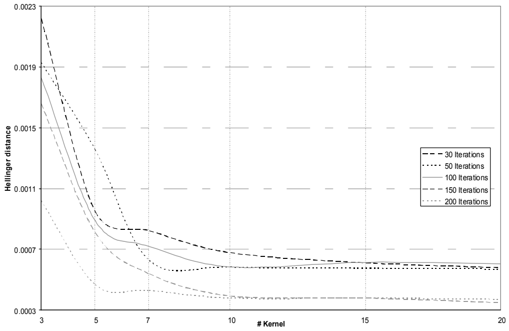

Figure 2. Distance average vs number of kernels, number of iterations as parameter

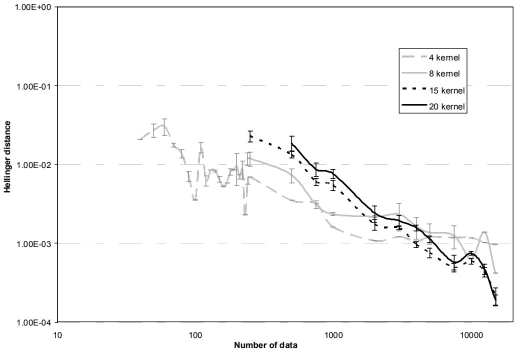

Figure 3. Distance average vs number of kernels, number of iterations as parameter

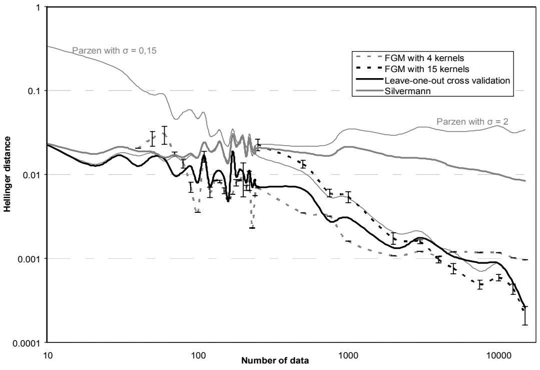

Figure 4. Comparison between the methods