Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/080236

The Effect Evaluation of Density Estimation through Non-Gaussian Measurement

Feng Zhao[1,2,*], *Zhiyong An*[1,2] *and Hailan Jiang* [3]

[1] School of Computer Science and Technology, Shandong Institute of Business and Technology, Yantai 264005, China
[2] Key Laboratory of Intelligent Information Processing in Universities of Shandong ,Shandong Institute of Business and Technology, Yantai 264005, China
[3] Information Engineering department, Shandong Polytechnic, Jinan 250104, China

Received: 8 Apr. 2013, Revised: 9 Aug. 2013, Accepted: 11 Aug. 2013
Published online: 1 Mar. 2014

Abstract: How to test the effect of density estimation methods is the key problem in the statistics. This paper presents a new criterion for assessing the effect of density estimation to select the suitable density estimation method, using the maximum-entropy non-Gaussian measurement. Comparing withχ^2-test and D_n-test, the method avoids the problem of the data interval division, and it is suitable for any type probability distribution. Simulation results show that the proposed method can accurately discriminate the pros and cons of different density estimation methods..

Keywords: Density estimation, hypothesis testing, non-Gaussian measurement

1 Introduction

Probability density function estimation is the key problem in the statistical learning, and we can solve almost all the problem based on the density function [1,2,3,4]. Many density estimation methods have been proposed. The common approach for density estimation is the parametric approach [5,6], such as maximum likelihood, Bayesian techniques, etc. The other one is the nonparametric approach [7,8], such as the kernel density estimator, the k-nearest neighbor technique and the neural networks, etc. Each method has its own merits and demerits for the different data sets. However, for the practical data, which kind of density estimation method is more effective? In other words, for the same data sets, which is the closest to the real one in the estimated density functions using the different density estimation methods?

At present, conventional hypothesis testing methods are divided into two categories[9]. The first one is the parametric test, which mainly is used to test the unknown parameters in the case that the distribution function form is known. However, it is difficult to know the distribution function form of a data set. The other one is the non-parametric test,such as χ^2-test and D_n-test, etc. However, the interval division is needed for χ^2-test, and D_n-test can only deal with the continuous distribution data set [9].Moreover, above methods only can be applied to verify whether a certain density estimation function $\widehat{f}(x)$ is suitable to the sampler data, and can't discriminate which is the better one between the density estimation functions $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$.

In this paper, using the maximum-entropy non-Gaussian measurement [10], we present a criterion for assessing the effect of density estimation methods, which can be applied to select the suitable density estimation method. Simulation results show that the method can compare different density estimation methods effectively and is suitable to any density distribution form data set.

2 The Test of Density Estimation Methods

2.1 Description of the Problem

Let $X = \{x_1, x_2, \cdots, x_n\}$denote a set of random sample. The underlying density is$y = f(x)$, and the distribution function is $F(x) = \int_{-\infty}^{x} f(t)dt$. Let and be the density functions which were estimated using the different density estimation methods. We need to discriminate which is near to $f(x)$ between $\widehat{f}_1(x)$ and $\widehat{f}_2(x)$?

* Corresponding author e-mail: zhaofeng1016@126.com

It is known that the distribution function $u = F(x)$ of a one-dimensional variable x is uniform in [0, 1][11,12]. We assume$X = \{x_1, x_2, \cdots, x_n\}$ is a set of random sample with the distribution function $X = \{x_1, x_2, \cdots, x_n\}$, therefore $u_i = F(x_i), (i = 1, 2, \cdots, n)$ can be seen as the sample from the uniform distribution in [0, 1]. Without loss of generality, we assume the points x_i are sorted in ascending order. Taking into account x_i is monotonically increasing, we have $u_1 \leq u_2 \leq \cdots \leq u_n$. When n is sufficiently large, u_i should be evenly spread in [0, 1]. Let $\widehat{f}(x)$ and $\widehat{f}(x)$ are the estimated density function and distribution function of the data sets X respectively. It is evidently that $\widehat{u}_i = \widehat{F}(x_i)$ should be more uniform scattered in [0, 1] if $\widehat{f}(x)$ is closer to $f(x)$.This means the degree of $\widehat{f}(x)$ approximation to $f(x)$ can be measured by the uniformity of $\widehat{u}_i = \widehat{F}(x_i)$. In the following, based on the non-Gaussian measurement, we give a method to assess the uniformity of $\widehat{u}_i = \widehat{F}(x_i)$.

In fact, according to the random number generator principle, $\widehat{u}_i = \widehat{F}(x_i)(i = 1, 2, \cdots, n)$ can generate n points $\widehat{y}_i$, which satisfies the standard Gaussian distribution. And the relationship of $\widehat{u}_i$ and $\widehat{y}_i$ can be written as follows

$$\widehat{y}_i = \Phi^{-1}(\widehat{u}_i) \tag{1}$$

where $u = \Phi(y)$ denotes the standard Gaussian distribution function and $y = \Phi^{-1}(u)$ is its inverse operation

Obviously, if $\widehat{y}_i = \Phi^{-1}(\widehat{u}_i)$ is more close to Gaussian distribution, u_i is more uniform in [0, 1]. Thereby, the degree of $\widehat{f}(x)$ approximation to $f(x)$ can be measured using the Gaussian degree of $\widehat{y}_i = \Phi^{-1}(\widehat{u}_i)$ approximation to the Gaussian distribution, that's to say, the gaussianity of $\widehat{y}_i = \Phi^{-1}(\widehat{u}_i)$.

2.2 The Gaussianity Measurement of One-Dimensional Random Variable

The gaussianity of a one-dimensional random variable can be measured by the maximum-entropy non-Gaussian measurement which is presented by Hyvarinen A [10]. It can be writen as follows

$$J(y) = [E\{G(y)\} - E\{G(v)\}]^2 \tag{2}$$

Where $v \sim N(0,1)$ is the standard Gaussian variable, and y is the random variable with the zero mean and unit variance, and G is a non-quadratic function 10. Evidently, if y is closer to the Gaussian distribution, $J(y)$ is smaller.

2.3 The Method Steps

Based on the above discussion, the main step is summarized as follows.

Step1. Let $\widehat{f}_j(x), (j = 1, 2, \cdots, m)$ is the estimated density function using the jth density estimation method, and $\widehat{F}_j(x)$ is the distrib ution function correspondingly.

Step2. Compute $\widehat{u}_j(x_i) = \widehat{F}_j(x_i)(i = 1, 2, \cdots, n)$ using the following formula.

$$\widehat{u}_j(x_i) = \widehat{F}_j(x_i)(i = 1, 2, \cdots, n) \tag{3}$$

Step3. Generate the random sample y_i^j which satisfies the standard Gaussian distribution according to $\widehat{u}_j(x_i)(i = 1, 2, \cdots, n)$(see Eq.(2)).

$$y_i^j = \Phi^{-1}(\widehat{u}_j(x_i)) \tag{4}$$

Step4. Compute the non-Gaussian measurement $J(y^j)$ using Eq. (1).

Step5. Identify the suitable density estimation method using the following rule.

$$j = \max_{j=1,2,\cdots,m} \{J(y^j)\} \tag{5}$$

3 Simulations

In order to evaluate the performance of the density estimation methods based on non-Gaussian measurement, two data sets X_1 and X_2 are generated using the Pseudo-random number generator. X_1 satisfies the Gaussian distribution $N(2,2)$, and X_2 uniform distribution (see Eq.(5)). Then, we estimate the density function of the above data sets using the parametric approach (Gaussian model) and the non-parametric approach (Parzen windows)[13]. Finally, we test the above density estimation functions for evaluating the performance of the proposed algorithm, using the method based on non-Gaussian measurement.

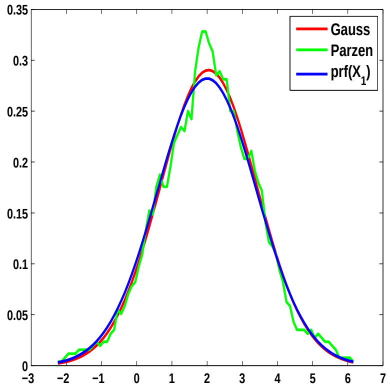
$$f(x) = \begin{cases} 1, \ -2.5 < x < -2 \\ 0.25, \ 0 < x < 2 \\ 0, \ other \end{cases} \tag{6}$$

The density estimation effect of Gaussian model and Parzen window method are shown in figure 1. Where the solid curve "—" denotes the probability density of the dates, and the long dash"−− " is the estimated one by Gaussian model, and the short dash"..." is the estimated one by Parzen window method. That shown in the table 1 is the non-Gaussian measurement and the mean square error of the true and estimated density function of Gaussian model and Parzen window method. Here, the formula of the mean square error can be obtained as

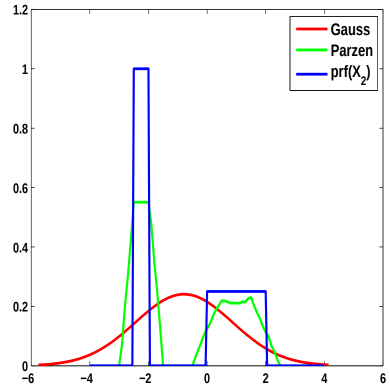
$$error = \frac{1}{n} \sum_{x \in X} \left| \widehat{f}(x) - f(x) \right|^2 \tag{7}$$

Where $f(x)$ and $\widehat{f}(x)$ are the true and estimated density functions, respectively; and n is the number of x.

As can be seen from Fig 1 and Table 1, first, for the data X_1 , the estimated density function can be better fitted to the true one using both Gaussian model and Parzen window method. Especially, the density curve of Gaussian model is more smooth and more close to the true distribution. The non-Gaussian measurements of them are almost close to zero. It means that the non-Gaussian measurement can discriminate the density estimation effect. Then, for the data X_2 whose distribution is the mixed uniform distribution, in comparison with Gaussian model, the result of Parzen window method is more close to the true density function. Correspondingly, the non-Gaussian measurement of Parzen window method is less than that of the Gaussian model. Lastly, it can be seen that the estimation effect of X_1 is superior to the estimation function of X_2 for the Gaussian model and Parzen window method.Similarly, the mean square error and the non-Gaussian measurement in X_1 is less than that in X_2.



(a) The Gaussian data X_1



(b) The uniform data X_2

Fig. 1: The density estimation effect

Table 1: Test of the density estimation effects

	The non-Gaussian measurements		the mean square error	
	Gaussian model	Parzen window	Gaussian model	Parzen window
X_1	$5.97 * 10^{-10}$	$6.41 * 10^{-5}$	$4.21 * 10^{-5}$	$3.97 * 10^{-4}$
X_2	0.007	0.0017	0.377	0.124

4 Conclusion

Based on the maximum-entropy non-Gaussian measurement, a criterion is presented for assessing the effect of density estimation methods. Comparing with the classical hypothesis testing methods, this method needs not divide the interval of the data set, and it is suitable to any density distribution form. The simulation results of two data sets show that the method is effective

Finally, the presented method in the paper is only suitable to one-dimensional density estimation. Further investigation is still needed for high dimensional density estimation.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grant No. 60970105 and 61173173, the Provincial Natural Science Foundation of Shandong under Grant No. ZR2011FM035 and ZR2011FL004, the Science and Technology plan project of Yan Tai under Grant No. 2011055. 1)the Key Project of Chinese Ministry of Education (2012101).

References

[1] J. Dai, S. Sperlich. Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation, Computational Statistics and Data Analysis, **54**, 2487-2497 (2010).
[2] K. T. Ting, J. R. Wells. Multi-dimensional mass estimation and mass-based clustering. IEEE International Conference on Data Mining, December 13-17; Sydney, Australia, (2010).
[3] J. S. Kim, C. D. Scott. L2 kernel classification, IEEE Transactions on Pattern Analysis and Machine Intelligence, **10**, 1822-1831 (2010).
[4] B. Han, D. Comaniciu, Y. Zhu, L. S. Davis. Sequential kernel density approximation and its application to real-time visual tracking, IEEE Transactions on Pattern Analysis and Machine Intelligence, **7**, 1186-1197 (2008).
[5] P. Stinis. A maximum likelihood algorithm for the estimation and renormalization of exponential densities, Computational Physics, **2**, 691-703 (2005).
[6] P. Lambert, P. H. C. Eilers. Bayesian density estimation from grouped continuous data, Computational Statistics and Data Analysis, **4**, 1388-1399 (2009).

[7] F. Comte, V. G. Catatot. Convolution power kernels for density estimation, Statistical Planning and Inference, **7**, 1698-1715 (2012).
[8] L. Elia, F. Corona, A. Lendasse. Residual variance estimation using a nearest neighbor statistic, Journal of Multivariate Analysis, **4**, 811-823 (2010).
[9] R. R. Wilcox. Introduction to Robust Estimation and Hypothesis Testing, Academic Press, Salt Lake City, USA, (2012).
[10] A. Hyv?rinen. New approximations of differential entropy for independent component analysis and projection pursuit. Proceedings of the conference on Advance in neural information processing systems. May 4-9, Anchorage, Alaska, USA, (1998).
[11] J. L. Devore. Probability and Statistics for Engineering and the Sciences. Duxbury press, CA, USA, (2011).
[12] M. I. Malik, A. Amir. Density estimation and random variate generation using multilayer networks, IEEE Transactions on Neural Networks, **3**, 497-530 (2002)
[13] Z. Q. Bian, X. G. Zhang. Pattern Recognition. Tsinghua University Press, Peiking, China, (2008).

Feng Zhao received the MS degree in computer science from Xi'Dian University in 2004, and the PhD degree in computer science from Xi'Dian University in 2008. He is currently a associate professor in Shandong Institute of Business and Technology. His research interests are in the areas of patter recognition and information processing.

Zhiyong An received the MS degree in computer science from Xi'Dian University in 2004, and the PhD degree in computer science from Xi'Dian University in 2008. He is currently a associate professor in Shandong Institute of Business and Technology. His research interests are in the areas of image retrieval and information processing.

Hailan Jiang received the master's degree in software engineering from Shandong University in 2004. She is currently an associate professor in Shandong Polytechnic. Her research interests are in the areas of information processing, mobile application development and Java web application