proceedings

MDPI

Extended Abstract

Nonparametric Mean Estimation for Big-but-Biased Data [†]

Laura Borrajo * and Ricardo Cao

Research Group MODES, CITIC, Department of Mathematics, University of A Coruña, 15071 A Coruña, Spain; ricardo.cao@udc.es

* Correspondence: laura.borrajo@udc.es; Tel.: +34-981-167-000-1301

† Presented at the XoveTIC Congress, A Coruña, Spain, 27–28 September 2018.

Published: 19 September 2018

Abstract: Some authors have recently warned about the risks of the sentence *with enough data, the numbers speak for themselves*. The problem of nonparametric statistical inference in big data under the presence of sampling bias is considered in this work. The mean estimation problem is studied in this setup, in a nonparametric framework, when the biasing weight function is unknown (realistic). The problem of ignoring the weight function is remedied by having a small SRS of the real population. This problem is related to nonparametric density estimation. The asymptotic expression for the MSE of the estimator proposed is considered. Some simulations illustrate the performance of the nonparametric method proposed in this work.

Keywords: Bias Correction; Big Data; Kernel Method; mean estimation; Nonparametric Inference

1. Introduction

At certain times a large sample is not representative of the population, but it is biased (B3D). Some of the problems coming from ignoring sampling bias in big data statistical analysis have been recently reported by Cao [1]. A good example cited by Crawford [2] is the data collected in the city of Boston through the StreetBump smartphone app that underestimates the number of potholes in some neighborhoods of the city, with the consequent deficient management of resources. Another example is the database of more than 20 million tweets generated by Hurricane Sandy. These data come from a biased sample of the population, since most of the tweets came from Manhattan, while few tweets were originated in the most affected areas by the catastrophe. In other examples, such as those cited in Hargittai [3], survey data show that the use of sites is biased yielding samples that limit the generalizability of findings.

In this context, let us consider a population with CDF F (density f) and consider a SRS, $\mathbf{X} = (X_1, \ldots, X_n)$, of size n from this population. Assume that we are not able to observe this sample but we observe, instead, another sample $\mathbf{Y} = (Y_1, \ldots, Y_N)$, of a much larger sample size ($N >> n$) from a biased distribution G (density g), such that $g(x) = w(x)f(x)$, for some weight function $w(x) \geq 0, \forall x$.

2. Mean Estimation in B3D

To deal with the mean estimation problem in this context, we propose the realistic estimator (unknown w case) whose motivation is explained by Cao and Borrajo [4]:

$$\hat{\mu}^{\hat{w}_{h,b}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\frac{Y_i}{\hat{w}_{h,b}\left(Y_i\right)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\hat{w}_{h,b}\left(Y_i\right)}} = \frac{\frac{1}{N}\sum_{i=1}^{N}Y_i\frac{\hat{f}_h\left(Y_i\right)}{\hat{g}_b\left(Y_i\right)}}{\frac{1}{N}\sum_{i=1}^{N}\frac{\hat{f}_h\left(Y_i\right)}{\hat{g}_b\left(Y_i\right)}}. \tag{1}$$

In order to work with this estimator, extra information is required. We propose a scenario in which, in addition to the biased sample, **Y**, we also observe a SRS, **X**, of small size of the real population. The Parzen-Rosenblatt KDE (see [5,6]) based on **X** and **Y** can be used to estimate f and g.

The final expression of the AMSE of (1) ($h \to 0$, $b \to 0$, $nh \to \infty$, $Nb \to \infty$ and $N/n \to \infty$) is:

$$\begin{aligned} AMSE\left(\hat{\mu}^{\hat{w}_{h,b}}\right) &= \left(C_1 b^2 + \frac{C_2}{Nb}\right)^2 + \frac{C_3}{n} + \frac{C_4}{Nn} + \frac{C_5}{N^2} + \frac{C_6}{Nnh} + \frac{C_7}{N^2 b} \\ &+ \frac{C_8 h^2}{N^2 b} + \frac{C_9 h^4}{N} + \frac{C_{10} b^4}{N} + \frac{C_{11} h^2 b^2}{N} + \frac{C_{12} h}{Nn} + \frac{C_{13} b}{N^2}. \end{aligned}$$

3. Case Study with Simulated Data

Let us consider $f(x) = \frac{3}{14}(x^2+1)\ \mathbf{1}_{[0,2]}(x)$ and $w(x) = 1.5\ \mathbf{1}_{[0,1.5]}(x) + x\ \mathbf{1}_{(1.5,2]}(x)$ (Figure 1a):

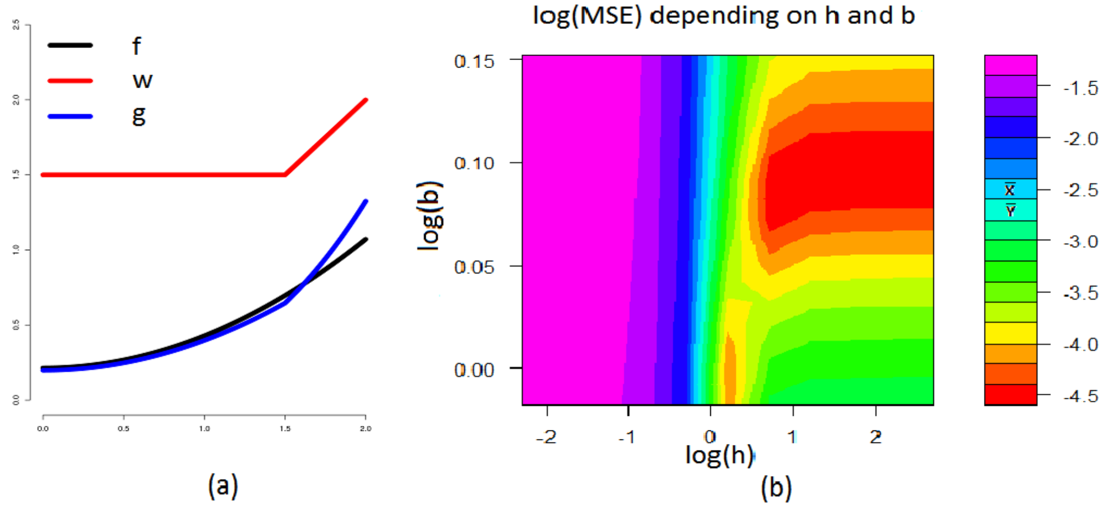

Figure 1. (**a**) Densities involved in the model. (**b**) Logarithm of the MSE of mu depending on the logarithm of h and b for this model, considering n = 100 and N = 10,000.

Figure 1b shows that the proposed estimator improves the estimation performed using the SRS, $\overline{X}$, and the biased sample, $\overline{Y}$, for a large number of combinations of h and b. Looking at Table 1, we observe that the best choice for h and b based on the simulation study contradicts the assumption ($h \to 0$, $b \to 0$) used in obtaining the asymptotic results. The AMSE for (1) under these non-standard asymptotic conditions ($h \to h_0$, $b \to b_0$) is:

$$AMSE\left(\hat{\mu}^{\hat{w}_{h_0,b_0}}\right) = \frac{D_1}{N} + \frac{D_2}{Nn} + \frac{D_3}{N^2} + \frac{D_4}{N^3}.$$

Table 1. MSE of the different estimators and optimal bandwidths obtained from the simulation study.

n	N	$MSE(\overline{X})$	$MSE(\overline{Y})$	$MSE(\hat{\mu}^{\hat{w}_{h,b}})$	h	b
10	100	2.9×10^{-2}	4.4×10^{-3}	2.4×10^{-3}	1.99	1.05
50	2500	5.6×10^{-3}	1.7×10^{-3}	9.9×10^{-5}	3.97	1.18
100	10,000	2.9×10^{-3}	1.6×10^{-3}	2.5×10^{-5}	5.00	1.20
500	250,000	5.0×10^{-4}	1.6×10^{-3}	1.1×10^{-6}	12.22	1.23
1000	1,000,000	2.0×10^{-4}	1.6×10^{-3}	2.7×10^{-7}	12.22	1.24

4. Conclusions

Big Data brings new statistical challenges since bias is much more present. Ideas from length-biased data and nonparametric smoothing techniques are important in this context, testing for bias is a relevant problem in Big Data and smoothing parameter selection may be paradoxical in B3D.

Funding: This research has been supported by MINECO Grants MTM2014-52876-R and MTM2017-82724-R and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the European Regional Development Fund (ERDF). The second author's research was sponsored by the Xunta de Galicia predoctoral grant (with reference ED481A-2016/367) for the universities of the Galician University System, public research organizations in Galicia and other entities of the Galician R&D&I System, whose funding comes from the European Social Fund (ESF) in 80% and in the remaining 20% from the General Secretary of Universities, belonging to the Ministry of Culture, Education and University Management of the Xunta de Galicia.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AMSE	Asymptotic mean squared error
B3D	Big-but-biased Data (BBBD)
CDF	Cumulative distribution function
KDE	Kernel density estimator
MSE	Mean squared error
SRS	Simple random sample

References

1. Cao, R. Inferencia estadística con datos de gran volumen. *Gac. RSME* **2015**, *18*, 393–417.
2. Crawford, K. The hidden biases in big data. *Harv. Bus. Rev.* **2013**. Available online: https://hbr.org/2013/04/the-hidden-biases-in-big-data (accessed on 4 April 2016).
3. Hargittai, E. Is Bigger Always Better? Potential Biases of Big Data Derived from Social Network Sites. *Ann. Am. Acad. Political Soc. Sci.* **2015**, *659*, 63–76.
4. Cao, R.; Borrajo, L. Nonparametric Mean Estimation for Big-But-Biased Data. In *The Mathematics of the Uncertain, Studies in Systems, Decision and Control*; Springer: Cham, Switzerland, 2018; Volume 142, pp. 55–65.
5. Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.
6. Rosenblatt, M. Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **1956**, *27*, 832–837.