Enliven Archive
Intensifying Thoughts

Editorial www.enlivenarchive.org Enliven: Biostatistics and Metrics

Nonparametric Kernel Methods with Errors-In-Variables: Advancements and Challenges

Kris De Brabanter*

Department of Statistics and Department of Computer Science, Iowa State University

*Corresponding author: Kris De Brabanter, Department of Statistics and Department of Computer Science, Iowa State University, Snedecor Hall, Ames, IA, 50010-1210, E-mail: kbrabant@iastate.edu

Received Date: 12th October 2015
Accepted Date: 15th October 2015
Published Date: 18th October 2015

Citation: De Brabanter K(2015) Nonparametric Kernel Methods With Errors-In-Variables: Advancements and Challenges. Enliven: Biostat Metr 2(1):00e1.

Abstract

In this brief manuscript we describe some of the recent advancements and challenges in the field of nonparametric regression with errors-in-variables. These problems arise (in principle) each time when data is measured due to the inaccuracy of the measuring equipment. We discuss 5 challenges in this field and provide recent references discussing this problem setting.

Keywords: Measurement errors; Errors-in-variables; Deconvolution

Introduction

In the classical statistical analyses it is often assumed the data is collected without measurement error or contamination. Simply stated, the random error is only assumed on the dependent variable and not on the independent variable. However, this is just a simplification of reality. Whenever data is measured in real life, there is always measurement error due to the inaccuracy of the measuring equipment, consider, for example, thermometers, chronographs, pressure sensors, etc. Taking this extra random error into account is usually referred to as a deconvolution problem or errors in-variables [1]. One of the earliest results considering this problem setting can be found in Berkson [2]. Further theoretical developments were done by Carroll and Devroye [3,4]. Deconvolution problems occur in many fields of nonparametric statistics, for example, density estimation based on contaminated data [5], nonparametric regression with errors-in-variables [6], image and signal deblurring [7]. During the last decades, these topics have received considerable attention. As applications of deconvolution procedures concern many real-life problems in econometrics, biometrics, medical statistics and image reconstruction. On the other hand, some rigorous results from Fourier analysis, functional analysis and probability theory are required to understand the construction of deconvolution techniques and their properties. Therefore making this field particularly interesting for mathematicians.

We will now introduce the mathematical concept of nonparametric regression with errors-in-variables. There are direct links with the density estimation setting with contaminated data. In standard nonparametric regression with errors-in-variables, we assume that the independent variables (covariates), X, can only be observed with some additive independent noise. Therefore, we change the observation scheme into the independent and identically distributed (i.i.d.) dataset $(W_1, Y_1), \ldots, (W_n, Y_n)$, where

$$W_j = X_j + \delta_j \text{ and } Y_j = m(X_j) + \varepsilon_j, \; j = 1, \ldots, n \quad (1)$$

where m is the regression function. The independent variable errors δ_j are i.i.d. unobservable random variables having error density g. Note that they are different from the regression errors ε_j. The δ_j are stochastically independent of the X_j and the Y_j. In order to proceed, we need one of the two following assumptions: the error density g is known or unknown. The optimal rates of convergence critically depend on this error density g, more specifically on the tail behavior of the Fourier transform of g. The following two types of error distributions are intensively studied in the deconvolution literature: ordinary smooth (e.g. Laplace, Gamma) and supersmooth (e.g. Cauchy, Gaussian) error densities. Simply said, densities characterized by the fact that their Fourier transforms decay in some finite power are called ordinary smooth. Densities whose Fourier transforms have exponential tails are called supersmooth. The supersmooth case turns out to be the hardest problem. We refer the interested reader to [8] for a rigorous theoretical study regarding deconvolution problems in nonparametric statistics.

Methods and Challenges

In case the independent variables are not affected by contamination i.e., $X_1, \ldots, X_n$ are directly observed, the Nadaraya-Watson (NW) kernel estimator with bandwidth h is a popular kernel regression estimator. However, in the measurement error case, we only observe contaminated data $W_1, \ldots, W_n$ and the Nadaraya-Watson estimator needs to be adapted. This modification was proposed by [9]. The NW estimator is a special case of the local polynomial kernel regression family [10]. One of their attractive features is their capacity to adapt automatically to boundary effects and hence reducing the bias with no or little variance increase (for certain polynomial orders). Extending the errors-in-variables case to the local polynomial framework was proposed by [11]. Their methodology consists of constructing unbiased estimators of the terms depending on the non-observable independent random variable X involved in the standard local polynomial regression estimators. The key to finding these unbiased estimators is the Fourier transform. This problem remained unsolved for more than 15 years!

Although serious theoretical progress is made regarding this complicated problem, some questions still remain (partially) unanswered:

- In practice we do not observe the contamination error δ. How can this be estimated?
- Kernel based regression estimators depend on a bandwidth h. Usually this is found by cross validation or plug-ins. Unfortunately they all depend on the fact that the independent variables $X_1, \ldots, X_n$ are directly observable.
- Does there exist an optimal deconvolution kernel? And if so, which one is it?
- How can this be extended to the multivariate case?
- How to implement this in a numerically stable way?

Earlier studies assumed a specific parametric form for the contamination density and estimated a specific parameter of that density without any additional observation [8,12]. However, in real life applications a known contamination density is a rather unrealistic assumption. A consistent estimator of m can only be constructed if the contamination density can be consistently estimated. Second, [13,14] assume that a sample of observations from the error density is available and estimate the density non parametrically from those data. A third approach, applicable when replicated measurements (panel data) are available, consists of estimating the contamination density from the replicates [15,16]. The latter being the most commonly used.

Most nonparametric regression estimators (not only kernel based ones) have one or more so-called tuning parameters. Usually, these are found via some data driven method such as cross-validation (CV). However, this method assumes that the independent variable, X, is directly observable and there does not seem to exist a straightforward extension of CV to the error-in-variables problem. Probably one of the first approaches to data-driven bandwidth selection, called SIMulation and EXtrapolation (SIMEX), in this setting was proposed by [17]. The key idea of this method is to generate new observations with increased noise levels. Then, fitting those data at different noise levels resulting in some appropriate estimated curve w.r.t. to the noise level in some real interval. An empirical version of the estimated quantity is then obtained by the value of the extrapolated curve at a noise level equal to zero. Delaigle and Hall [17] suggested to add some additional independent noise to the independent variables. Their numerical simulations indicate that this methods tends to work well for this setting.

In general, there are close parallels between the optimal kernel choice in nonparametric density deconvolution and its counterpart in density estimation. However, certain aspects of these problems are strikingly different. Therefore, Delaigle and Hall [5] stated the following: "this property leads us to conclude that optimal kernels do not give satisfactory performance when applied to deconvolution." The reason for this is that certain standard side conditions are necessary in this setting. However, at the time of writing, we are not aware of a similar study being performed for the nonparametric regression case with errors-in-variables. However, looking at the exact mean integrating squared error expression suggests that kernels which have a flat top in the Fourier domain are quite suitable [1].

Finally, most authors study the univariate regression case but many real data sets are bivariate or multivariate. Although there exist an extensive literature on multivariate density deconvolution, the multivariate regression case is severely lacking. Fan and Masry [18] and Masry [19] established asymptotic properties and considered the case of stationary random processes with errors-in-variables respectively. But no data-driven procedure for bandwidth selection was suggested. Besides the theoretical aspects, the numerical implementation of the deconvolution framework is equally important. In general, computing these estimators requires special care. Since there are no closed form solutions at hand for the integrals involved in the calculations numerical techniques have to be used. It is known that the integrand can oscillate quite severely causing fast numerical integration algorithms to fail. This problem can be avoided by using the fast Fourier transform [20,21]. There exist an R package decon [22] which implements the deconvolution estimator for regression (local constant regression) and provides some bandwidth selection criteria. Unfortunately, as reported in [23] there are some problems with this package (version 1.2-4).

Conclusion

In this brief manuscript we gave an overview of advancements and challenges in deconvolution problems. Classical kernel regression estimation (and also density estimation) suffer when the independent variables are contaminated with measurement error. This is due to the fact the classical estimators inherently assume an error-free independent variable. In order to allow that the estimator can deal with these measurement errors or error-in-variables, the classical estimators need to be modified. The key method in the deconvolution approach is the Fourier transform.

References

1. Meister A (2009) Deconvolution Problems in Nonparametric Statistics. Springer 193.
2. Berkson J (1950) Are there two regressions? J Am Stat Assoc 45: 164-180.
3. Carroll RJ, Hall P (1988) Optimal rates of convergence for deconvolving a density. J Am Stat Assoc 83: 1184-1186.
4. Devroye L (1989) Consistent deconvolution in density estimation. Can J Stat 17: 235-239.
5. Delaigle A, Gijbels I (2006) Data-driven boundary estimation in deconvolution problems. Computational Statistics & Data Analysis 50: 1965-1994.

6. Delaigle A, Meister A (2011) Rate-optimal nonparametric estimation in classical and Berkson errors-in-variables problems. J Stat Plan Inference 141: 102-114.
7. Qiu P (2005) Image Processing and Jump Regression Analysis. John Wiley & Sons.
8. Meister A (2006) Density estimation with normal measurement error with unknown variance. Statistica Sinica 16: 195-211.
9. Fan J, Truong YK (1993) Multivariate regression estimation with errors-in-variables: Asymptotic normality for mixing processes. J Multivar Anal 43: 237-271.
10. Fan J, Gijbels I (1996) Local Polynomial Modelling and Its Applications. Chapman & Hall.
11. Delaigle A, Gijbels I (2006) On optimal kernel choice for deconvolution. Statistics & Probaility Letters 76: 1594-1602.
12. Butucea C, Matias C (2005) Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. Bernoulli 11: 309-340.
13. Diggle P, Hall P (1993) A Fourier approach to non-parametric deconvolution of a density estimate. J R Stat Soc Series B Stat Methodol 55: 523-531.
14. Neumann MH (1997) On the effect of estimating the error density in nonparametric deconvolution. J Nonparametr Stat 7: 307-330.
15. Horowitz JL, Markatou M (1996) Semiparametric estimation of regression models for panel data. Rev Econ Stud 63: 145-168.
16. Delaigle A, Fan J, Carroll RJ (2009) A design adaptive local polynomial estimator for the errors-in-variables problem. J Am Stat Assoc 104: 348-359.
17. Delaigle A, Hall P (2008) Using SIMEX for smoothing-parameter choice in errors-in-variables problems. J Am Stat Assoc 103: 280-287. 18. Fan J, Masry E (1992) Nonparametric regression with errors in variables. Ann Stat 21: 1900-1925.
18. Fan J, Masry E (1992) Nonparametric regression with errors in variables. Ann Stat 21: 1900-1925.
19. Masry E (1993) Multivariate regression estimation with errors-in-variables for stationary processes. Journal of Nonparametric Regression 3: 13-36.
20. Bailey DH, Swarztrauber PN (1994) A fast method for the numerical evaluation of continuous Fourier and Laplace transforms. SIAM J Sci Comput 15: 1105-1110.
21. Inverarity GW (2002) Fast computation of multidimensional Fourier integrals. SIAM J Sci Comput 24: 645-651.
22. Wang XF, Wang B (2011) Deconvolution Estimation in Measurement Error Models: The R Package decon. J Stat Softw 39.
23. Delaigle A (2014) Nonparametric kernel methods with errors-in-variables: Constructing estimators, computing them, and avoiding common mistakes. Australian & New Zealand Journal of Statistics 56: 105-124.