Journal of Multivariate Analysis 133 (2015) 38–50

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

Conditional density estimation in measurement error problems

Xiao-Feng Wang [a,*], Deping Ye [b]

[a] Department of Quantitative Health Sciences / Biostatistics Section, Cleveland Clinic Lerner Research Institute, Cleveland, OH 44195, USA

[b] Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL A1C 5S7, Canada

ARTICLE INFO

Article history:
Received 21 March 2013
Available online 18 September 2014

AMS subject classifications:
62G07
62G20

Keywords:
Measurement error
Gene microarray
Conditional density
Deconvolution
Ridge parameter
Kernel
Bandwidth selection

ABSTRACT

This paper is motivated by a wide range of background correction problems in gene array data analysis, where the raw gene expression intensities are measured with error. Estimating a conditional density function from the contaminated expression data is a key aspect of statistical inference and visualization in these studies. We propose re-weighted deconvolution kernel methods to estimate the conditional density function in an additive error model, when the error distribution is known as well as when it is unknown. Theoretical properties of the proposed estimators are investigated with respect to the mean absolute error from a "double asymptotic" view. Practical rules are developed for the selection of smoothing-parameters. Simulated examples and an application to an Illumina bead microarray study are presented to illustrate the viability of the methods.

1. Introduction

Measurement error problems have attracted a great deal of interest in the past two decades. A variety of models and methods for the problems have been applied in scientific fields, such as medicine, economy, and astronomy. Statistical deconvolution is an important component in measurement error models. The fundamental objective of deconvolution is to recover the unknown probability density function of a random variable when its observed values are contaminated with error. Let X be the variable of interest, which cannot be observed directly. Instead, we observe a sample of W,

$$W_j = X_j + U_j, \quad \text{for } 1 \le j \le n, \tag{1}$$

where X_j's are identically distributed as X, U_j's are identically distributed as U, and they are totally independent. The most popular approach to estimate the density of X is the *deconvolution kernel estimator* through applying an inverse Fourier transform and a kernel technique [3,31,14,15]. Other estimation procedures include the truncated Fourier inversion method [11], the wavelet-based method [17], the penalization approach [4], among others. Deconvolution problems based on more complicated model settings have also been extensively studied. Delaigle and Meister [8], Wang et al. [32], and McIntyre and Stefanski [26] considered the problems of heteroscedastic measurement errors. Hall and Maiti [19] investigated nonparametric deconvolution methods in two-level mixed models. Neumann [28], Johannes [25] and Wang and Ye [34] studied the

* Correspondence to: Department of Quantitative Health Sciences, Cleveland Clinic Lerner Research Institute, 9500 Euclid Avenue/JJN3, Cleveland, OH 44195, USA.

E-mail addresses: xfwang@gmail.com, wangx6@ccf.org (X.-F. Wang).

http://dx.doi.org/10.1016/j.jmva.2014.08.011
0047-259X/

density deconvolution with unknown error distribution. Delaigle and Meister [9] investigated kernel deconvolution when the characteristic function of the measurement errors contains zeros. Wang and Wang [33] discussed fast Fourier transform algorithms in measurement error models and developed an R software package. The literature on deconvolution problems is particularly large and is surveyed in the monograph by Meister [27].

In this paper, we consider the estimation problem of the conditional density of X given W, $f_{X|W}$, from the contaminated data W_j's. The problem is motivated by a wide range of background correction problems in gene array data analysis. Gene microarray techniques have become very popular in medical studies. A microarray is a collection of microscopic DNA spots attached to a solid surface. Hundreds of thousands of gene expression values are obtained from one array chip simultaneously. However, reading the expression values from a microarray is a noisy measurement process. The sources of measurement error include, for instance, irregularities in the array surface, variations in the laboratory process, different image scanner settings, and dye effects.

Typically, the first step in gene array data analysis is known as *background correction*, which refers to adjustments to the contaminated data intended to remove measurement error from the measured signal. Estimating the conditional density function from contaminated gene expression data, therefore, plays a key aspect of statistical inference and visualization here. It provides the most informative summary of the relationship between the contaminated gene intensities and the unobserved true signals. The current popular model of background correction in bioinformatics is the *normal-exponential* model [24,30]. It assumes that the observed intensity is equal to the true intensity plus the background noise, where the true signal follows an exponential distribution with mean α, and the background noise follows a normal distribution with mean μ and variance σ^2. However, the validity of the parametric assumptions is unknown in real gene array studies. Thus, it is of particular interest to nonparametrically estimate the conditional density from the contaminated gene intensities.

A variety of papers discuss the nonparametric conditional density estimation when bivariate data are available. Hyndman et al. [23] studied a kernel estimator. Bashtannyk and Hyndman [1] and Fan and Yim [18] proposed several rules for selecting smoothing parameters. De Gooijer and Zerom [5] proposed a modification of the Nadaraya–Watson type of smoother. Hall et al. [20] discussed cross-validation and the estimation of conditional probability densities. Efromovich [13] studied the conditional density estimation in a regression setting.

Unlike the conventional conditional density estimation problem from bivariate data, the observations for the variable-of-interest, X, are not available in the measurement error problem. In this paper, we investigate the estimation of the conditional density $f_{X|W}$ from the only-available contaminated sample W_j's under the model (1). In Section 2, estimators of $f_{X|W}$ are constructed in case of a known and an unknown error density. In Section 3, theoretical properties of the estimators are investigated with respect to the mean absolute error. In Section 4, practical rules are developed for the selection of smoothing-parameters. Simulated examples and an application to an Illumina bead microarray study are presented in Section 5. The proofs of theorems are given in the Appendix and some additional asymptotic results are provided in the supplement of the article (see Appendix B).

2. Methodology

Under the additive measurement error model (1), let f_X, f_U, and f_W be the density functions of X, U, and W, respectively. Denote $f_{X,W}(x, w)$ as the joint density of (X, W). The conditional density of X given $W = w$ is

$$f_{X|W}(x|w) = f_{X,W}(x, w)/f_W(w) = f_U(w - x)f_X(x)/f_W(w). \tag{2}$$

2.1. Estimation of $f_{X|W}$ with known error distribution

If one assumes that the error density f_U is known explicitly, f_X can be estimated by the classical deconvolution kernel approach. It is given by,

$$\hat{f}_X(x) = \frac{1}{n}\sum_{j=1}^{n} K_h^*\left(x - W_j\right), \tag{3}$$

where $K_h^*(\cdot) = K^*(\cdot/h)/h$,

$$K^*(z) = \frac{1}{2\pi}\int e^{-itz}\frac{\phi_K(t)}{\phi_U(t/h)}dt, \tag{4}$$

is known as the *deconvoluting kernel*, and $h > 0$ is a smoothing parameter. In (4), ϕ_U is the characteristic function of U, and $\phi_K(t) = \int e^{itx}K(x)dx$ is the Fourier transform of $K(x)$, a symmetric probability kernel with a finite variance $\int x^2K(x)dx < \infty$. Under the common assumption that ϕ_K is compactly supported and ϕ_U does not vanish on the real line, the deconvoluting kernel $K^*(\cdot)$ is well defined and finite.