Journal of the Korean Statistical Society 44 (2015) 187–201

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Least squares sieve estimation of mixture distributions with boundary effects

Mihee Lee [a], Ling Wang [b,*], Haipeng Shen [b], Peter Hall [a], Guang Guo [c], J.S. Marron [b]

[a] Department of Mathematics and Statistics, University of Melbourne Parkville, VIC, Australia

[b] Department of Statistics and Operations Research, University of North Carolina Chapel Hill, NC 27599, USA

[c] Department of Sociology, University of North Carolina Chapel Hill, NC 27599, USA

ARTICLE INFO

Article history:
Received 30 November 2013
Accepted 15 July 2014
Available online 1 August 2014

AMS 2000 subject classifications:
primary 62G08
secondary 65F30

Keywords:
Least squares
Measurement error
Mixture distribution
Penalization
Sieve method

ABSTRACT

In this study, we propose two types of sieve estimators, based on least squares (LS), for probability distributions that are mixtures of a finite number of discrete atoms and a continuous distribution under the framework of measurement error models. This research is motivated by the maximum likelihood (ML) sieve estimator developed in Lee et al. (2013). We obtain two types of LS sieve estimators through minimizing the distance between the empirical distribution/characteristic functions and the model distribution/characteristic functions. The LS estimators outperform the ML sieve estimator in several aspects: (1) they need much less computational time; (2) they give smaller integrated mean squared error; (3) the characteristic function based LS estimator is more robust against mis-specification of the error distribution. We also use roughness penalization to improve the smoothness of the resulting estimators and reduce the estimation variance. As an application of our proposed LS estimators, we use the Framingham Heart Study data to investigate the distribution of genetic effects on body mass index. Finally asymptotic properties of the LS estimators are investigated.

1. Introduction

In this paper, we consider measurement error models where we observe only the error-contaminated variable $Y = X + Z$, where X is the unobservable random variable of interest, and Z is the measurement error with a known density f_z that is independent of X. We are interested in estimating the distribution of X, which is assumed to be a mixture of several point masses and a continuous distribution. We are particularly interested in the case that the continuous part is supported on a finite interval, and has non-smooth boundaries.

Distribution estimation in measurement error models has been widely studied, but most of the earlier studies focused on estimating continuous density functions. Recently there are two studies (Lee, Shen, Burch, & Marron, 2010; van Es, Gugushvili, & Spreij, 2008) which consider mixtures of one discrete atom and one continuous component in the context of measurement error models, and independently propose the same estimator. The convergence rate of the estimator is recently derived by Gugushvili, Van Es, and Spreij (2011).

[*] Corresponding author.
E-mail addresses: miheel@unimelb.edu.au (M. Lee), lingw@email.unc.edu (L. Wang), haipeng@email.unc.edu (H. Shen), halpstat@ms.unimelb.edu.au (P. Hall), guang_guo@unc.edu (G. Guo), marron@unc.edu (J.S. Marron).

http://dx.doi.org/10.1016/j.jkss.2014.07.003

In terms of purely continuous distributions, there are two major types of deconvolution approaches. The first type uses ideas of Fourier and inverse Fourier transformation along with nonparametric smoothing. For example, see Lee et al. (2010), van Es et al. (2008) and references therein. The second type includes non-Fourier based deconvolution methods. In this group, many studies first employ basis functions such as B-splines or wavelets to expand the target density (or distribution) function, and then estimate the basis coefficients using various approaches. The studies include Johnstone, Kerkyacharian, Picard, and Raimondo (2004), Staudenmayer, Ruppert, and Buonaccorsi (2008) and references therein. In addition, several alternatives for deconvolution have been proposed, such as NPMLE, SIMEX, and TAYLEX, which are well reviewed in Carroll, Ruppert, Stefanski, and Crainiceanu (2006), Wagner and Stadtmüller (2008), and Wang, Sun, and Fan (2009).

Compared with the other studies, Lee et al. (2013) covers more general cases of measurement error models that have two features: (1) discrete and continuous mixtures and (2) non-smooth boundaries. First they approximate the distribution of X using discretization, which gives a *sieve* of the distribution family. Then they estimate the distribution using maximum likelihood (ML) within each sieve. Sieve type estimators have been proposed for deconvolution problems by Cordy and Thomas (1997) where degenerate distributions are used to approximate the continuous mixture component. In the error-free case, Ruppert, Nettleton, and Hwang (2007) proposed a sieve type density estimator for certain special distributions with known boundaries.

However, the ML method of Lee et al. (2013) involves long computation time and is not robust against misspecified error distribution. In this study, we propose alternative least squares (LS) sieve estimators based on the cumulative distribution function and characteristic function, instead of maximum likelihood. Our simulation results clearly demonstrate the advantages of the LS estimators. First, computational cost is much smaller when using the LS method. For example, it takes 113.32 s for the ML method on a simulated data set with sample size of 329, while it only takes 0.30 s using the cumulative distribution function based LS estimator. Secondly, in Section 4.1 the LS estimators are seen to give smaller (integrated) mean squared error. Furthermore, as seen in Section 4.2, the LS estimators are more robust when the error distribution is misspecified.

The remainder of the paper is organized as follows. Section 2 explicitly describes our model, and then proposes the two LS-sieve estimators, along with their estimation algorithms. In Section 3, consistency of the proposed estimators is established under appropriate regularity assumptions. Section 4 illustrates numerical performance of the various methods via simulation studies, and compares the ML-sieve with the LS-sieve estimators. Section 5 contains an application to the Framingham Heart Study data. Our methods are used to identify the distribution of some important SNPs' effects on body mass index (BMI). We conclude the paper in Section 6 with discussion of future work. Technical proofs are provided in the Appendix.

2. The model and the estimators

2.1. The model

Suppose that we can only observe an error contaminated variable Y, instead of X whose distribution is a mixture of several point masses plus a continuous distribution. That is,

$$Y = X + Z, \tag{1}$$

where Z is a measurement error with known density f_Z, and is independent of X. Our goal is to use a random sample $Y_1, \ldots, Y_n$ to estimate f_X, the generalized density of X, which is a mixture of discrete point masses $a_l,\ l = 1, \ldots, \nu$, and a continuous random variable X_c with density f_c, using weights $\pi_1, \ldots, \pi_\nu$, and $\pi_{\nu+1}$. Hence, the generalized density $f_X(x)$ has the following form:

$$f_X(x) = \sum_{l=1}^{\nu} \pi_l \delta_{a_l}(x) + \pi_{\nu+1} f_c(x), \tag{2}$$

where δ_{a_l} is the Dirac delta function at a_l. Here, the weights are probabilities in the sense that each π_l is nonnegative, $\sum_{l=1}^{\nu+1} \pi_l = 1$. We are particularly interested in the case where f_c is supported on a finite interval $[a, b]$. This paper focuses on scenarios where the values ν and $a_1, \ldots, a_\nu$ are known. In this setting, the estimation of f_X is equivalent to the estimation of both f_c and $\pi = (\pi_1, \ldots, \pi_{\nu+1})^T$. However, limited empirical results suggest that our method can be extended to cases where the locations of the point mass are unknown, which we will discuss in Section 6.

The first step is discretization of the continuous variable X_c. We approximate X_c by a discrete random variable $\tilde{X}_c$ taking values on an equally spaced grid, with grid spacing h. The discrete variable $\tilde{X}_c$ takes on values $x_j : x_{j+1} - x_j = h, j = 1, \ldots, r$, which cover the support of f_c. In practice, we choose $\tilde{X}_c$ satisfying

$$\tilde{X}_c = x_j \quad \text{if and only if } X_c \in [x_j - 0.5h, x_j + 0.5h).$$

The parameter h plays a role similar to the bin width in histogram estimation, and the same as the smoothing parameter in kernel density estimation. Let $\theta = (\theta_1, \ldots, \theta_r)^T$ be the probability distribution of $\tilde{X}_c$, i.e.

$$\theta_j = P(\tilde{X}_c = x_j) \quad \text{for each } j = 1, \ldots, r,$$

where $\theta_j \geq 0$ and $\sum \theta_j = 1$. Then each θ_j approximates the probability that X_c lies in the interval $[x_j - 0.5h, x_j + 0.5h)$.