Statistics and Probability Letters 123 (2017) 171–176

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Deconvolution of $\mathbb{P}(X < Y)$ with compactly supported error densities

Dang Duc Trong [a], Ton That Quang Nguyen [b], Cao Xuan Phuong [c,*]

[a] Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh National University, No. 227 Nguyen Van Cu Street, Ward 4, District 5, Ho Chi Minh City, Viet Nam

[b] Faculty of Fundamental Science, Industrial University of Ho Chi Minh City, No. 12 Nguyen Van Bao Street, Ward 4, Go Vap District, Ho Chi Minh City, Viet Nam

[c] Faculty of Mathematics and Statistics, Ton Duc Thang University, No. 19 Nguyen Huu Tho Street, Tan Phong Ward, District 7, Ho Chi Minh City, Viet Nam

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 20 December 2016
Accepted 20 December 2016
Available online 28 December 2016

MSC:
62G05
62G99

Keywords:
Deconvolution
Estimator
Convergence rate
Compactly supported

ABSTRACT

We study the problem of estimating the probability $\mathbb{P}(X < Y)$ when two random variables X and Y are observed with additional errors. We derive the convergence rate for a proposed estimator of $\mathbb{P}(X < Y)$ in the case of compactly supported error densities.

1. Introduction

In this paper, we study the problem of estimating the probability

$$\theta := \mathbb{P}(X < Y), \tag{1}$$

where the quantities X and Y are two independent random variables with unknown density functions f_X and f_Y, respectively. This problem arises from some studies in applied sciences, such as reliability and medicine. In stress–strength models of reliability theory (see, e.g. Kotz et al., 2003), the random variables Y and X represent strength of a component and stress impacting on the component, respectively. If the stress overcomes the strength, say $X > Y$, the component fails. Then the probability θ is defined as reliability of the component. In medicine (see, e.g. Zhou, 2008), the quantity θ is related to a receiver operating characteristic (ROC) curve which is used as a graphical tool for assessing the accuracy of a diagnostic test. More concretely, if X and Y represent continuous-scale diagnostic test measurements for a non-diseased and diseased patient, respectively, the ROC curve can be defined as the graph of the function $R(t) := 1 - F_Y\left(1 - F_X^{-1}(1-t)\right), 0 < t < 1$, where F_X and F_Y are the distribution functions of X and Y, respectively. In that case, θ is the area under the graph.

[*] Corresponding author.
E-mail addresses: ddtrong@hcmus.edu.vn (D.D. Trong), tonthatquangnguyen@iuh.edu.vn (T.T.Q. Nguyen), caoxuanphuong@tdt.edu.vn (C.X. Phuong).

http://dx.doi.org/10.1016/j.spl.2016.12.014
0167-7152/© 2016 Elsevier B.V. All rights reserved.

Motivated partly by the applications, the problem has been widely considered in the statistical literature (see, e.g. Mann and Whitney, 1947; Sen, 1967; Reiser and Guttman, 1986; Kundu and Gupta, 2006; Zhou, 2008; Montoya and Rubio, 2014, among others). However, most of the research has focused on estimating θ on the basis of the direct data $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} f_X$ and $Y_1, \dots, Y_m \overset{\text{i.i.d.}}{\sim} f_Y$. In many situations, the direct data are not available due to measurement errors. An example of the mentioned measurements was provided in Shear et al. (1987). The authors gave the noise data of systolic and diastolic blood pressure which have been used to predict future hypertension of children. Both because of the instrument used and the person recording the measurements, the data are highly susceptible to measurement errors. As a result, we only observe two independent random samples of $X' = X + \zeta$ and $Y' = Y + \eta$, where ζ and η are considered as additional errors. Therefore, we have naturally the problem of estimating θ from the samples of X' and Y'. In the paper, suppose that we have observations $X'_1, \dots, X'_n \overset{\text{i.i.d.}}{\sim} f_{X'}$ and $Y'^{1}, \dots, Y'^{m} \overset{\text{i.i.d.}}{\sim} f_{Y'}$, where

$$X_{j'} = X_j + \zeta_j, \qquad Y_{k'} = Y_k + \eta_k, \quad j = 1, \dots, n, \; k = 1, \dots, m. \tag{2}$$

Here, we assume that X_j, $\zeta_{j'}$, Y_k, $\eta_{k'}$ are mutually independent for $1 \leqslant j, j' \leqslant n$, $1 \leqslant k, k' \leqslant m$. In addition, the random variables ζ_j and η_k are i.i.d. with known densities f_ζ and f_η, respectively. The densities f_ζ, f_η are called error densities.

In a nonparametric setting, a classical estimator of θ (see, e.g. DeLong et al., 1988) is the Wilcoxon–Mann–Whitney statistic $\hat{\theta}^{\mathrm{WMW}} := (nm)^{-1} \sum_{j=1}^{n} \sum_{k=1}^{m} \mathbb{I}_{\left\{X'_j < Y'_k\right\}}$ where $\mathbb{I}_A$ is the indicator function of the set A. The bias of $\hat{\theta}^{\mathrm{WMW}}$ was studied in Coffin and Sukhatme (1997). Kim and Gleser (2000) proposed an estimation procedure for θ based on the SIMEX method when f_ζ, f_η are standard normal. Recently, Dattner (2013) also considered the problem when f_ζ, f_η are supersmooth densities. He proposed an estimator of θ of the form

$$\hat{\theta}^D = \frac{1}{2} - \frac{1}{nm} \sum_{j=1}^{n} \sum_{k=1}^{m} \frac{1}{\pi} \int_0^U \frac{1}{t} \operatorname{Im} \left\{ \frac{e^{it\left(X'_j - Y'_k\right)}}{f_\zeta^{\mathrm{ft}}(t) \overline{f_\eta^{\mathrm{ft}}(t)}} \right\} dt, \tag{3}$$

where U is a regularization parameter that must be chosen, and the notation h^{ft} denotes the Fourier transform of h. Dattner derived the optimal convergence rate uniformly on the Sobolev class with respect to the error $(\mathbb{E}|\hat{\theta}^D - \theta|^2)^{1/2}$. However, the estimator $\hat{\theta}^D$ is only defined if f_ζ^{ft} and f_η^{ft} are non-vanishing on $\mathbb{R}$. This will be violated if f_ζ and f_η are uniform densities or compactly supported densities in general. To the best of our knowledge, the problem of estimating θ in which f_ζ, f_η are compactly supported densities has still not been studied, so this case will be considered in the present paper.

For convenience, we introduce some notations. The convolution of two functions f and g is denoted by $f * g$. The Fourier transform of a density h is the function $h^{\mathrm{ft}}(t) = \int_{-\infty}^{\infty} e^{itx} h(x)\, dx$, $t \in \mathbb{R}$. For $\phi \in L^2(\mathbb{R})$, the $L^2(\mathbb{R})$-norm of ϕ is denoted by $\|\phi\|_2$. For a complex number z, the numbers $\operatorname{Re}\{z\}$, $\operatorname{Im}\{z\}$ and $\bar{z}$ denote the real part, the imaginary part and the conjugate of z, respectively. For two sequences of real numbers (a_n) and (b_n), the notation $a_n \leqslant O(b_n)$ means $a_n \leqslant \text{const} \cdot b_n$ for large n. The number $\lambda(A)$ is the Lebesgue measure of a measurable set $A \subset \mathbb{R}$. Finally, for a function $\varphi : \mathbb{R} \to \mathbb{R}$, the notation $\operatorname{supp}(\varphi)$ denotes support of φ, the closure in $\mathbb{R}$ of the set $\{x \in \mathbb{R} : \varphi(x) \neq 0\}$.

The rest of our paper is organized as follows. In Section 2, we derive our estimator. In Section 3, we provide some results on convergence rate and provide a confidence interval for θ. Section 4 presents numerical results based on some simulations. Proofs of all results in Section 3 are given in the supplementary paper (Trong et al., 2016).

2. Estimator

Let $Z = X - Y$. Then from (1) we obtain $\theta = \mathbb{P}(Z < 0) = F_Z(0)$, where F_Z is the continuous distribution function of Z. As is well known,

$$F_Z(x) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{1}{t} \operatorname{Im}\left\{e^{-itx} f_Z^{\mathrm{ft}}(t)\right\} dt, \quad x \in \mathbb{R},$$

where f_Z is the density function of the random variable Z and $i = \sqrt{-1}$. Since $f_Z^{\mathrm{ft}} = f_X^{\mathrm{ft}} \cdot \overline{f_Y^{\mathrm{ft}}}$, we have

$$\theta = F_Z(0) = \frac{1}{2} - \frac{1}{\pi} \int_0^\infty \frac{1}{t} \operatorname{Im}\left\{f_X^{\mathrm{ft}}(t) \overline{f_Y^{\mathrm{ft}}(t)}\right\} dt. \tag{4}$$

From (4), to obtain an estimator of θ, we will first try to construct suitable estimators of $f_X^{\mathrm{ft}}(t)$ and $f_Y^{\mathrm{ft}}(t)$. To this end, we will use the method of Tikhonov regularization. Consider the linear operator $A(\phi) = \phi f_\zeta^{\mathrm{ft}}$, $\phi \in L^2(\mathbb{R})$. For a real number $a > 1$, let W_a be the set of all functions $\phi \in L^2(\mathbb{R})$ satisfying $\int_{-\infty}^{\infty} |\phi(t)|^2 |t|^a\, dt < \infty$. Suppose that $f_X^{\mathrm{ft}} \in W_a$. For every $\delta > 0$, we define the Tikhonov functional $J_\delta(\phi) := \left\|A(\phi) - f_{X'}^{\mathrm{ft}}\right\|_2^2 + \delta \|\phi\, |I(\cdot)|^{a/2}\|_2^2$, $\phi \in W_a$, where $I(t) := t$. Then the minimizer ϕ_δ of J_δ satisfies the equation $(A^*A)(\phi_\delta) + \delta\, |I(\cdot)|^a \phi_\delta = A^*\left(f_{X'}^{\mathrm{ft}}\right)$, where A^* is the adjoint operator of A. Since $A^*(\varphi) = \varphi \overline{f_\zeta^{\mathrm{ft}}}$, $\varphi \in L^2(\mathbb{R})$, we obtain from the equation that $\phi_\delta\left(|f_\zeta^{\mathrm{ft}}|^2 + \delta\, |I(\cdot)|^a\right) = f_{X'}^{\mathrm{ft}} \overline{f_\zeta^{\mathrm{ft}}}$,