Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures‡

Albert Vexler, **Enrique F. Schisterman**[*,†], and **Aiyi Liu**
National Institute of Child Health and Human Development, U.S.A.

SUMMARY

Additive measurement errors and pooling design are objectively two different issues, which have been separately and extensively dealt with in the biostatistics literature. However, these topics usually correspond to problems of reconstructing a summand's distribution of the biomarker by the distribution of the convoluted observations. Thus, we associate the two issues into one stated problem. The integrated approach creates an opportunity to investigate new fields, e.g. a subject of pooling errors, issues regarding pooled data affected by measurement errors. To be specific, we consider the stated problem in the context of the receiver operating characteristic (ROC) curves analysis, which is the well-accepted tool for evaluating the ability of a biomarker to discriminate between two populations. The present paper considers a wide family of biospecimen distributions. In addition, applied assumptions, which are related to distribution functions of biomarkers, are mainly conditioned by the reconstructing problem. We propose and examine maximum likelihood techniques based on the following data: a biomarker with measurement error; pooled samples; and pooled samples with measurement error. The obtained methods are illustrated by applications to real data studies.

Keywords

deconvolution; design of experiments; Fourier inversion; infinitely divisible distribution; measurement error; pooling blood samples; receiver operating characteristic curves; stable distribution; summand's distribution

1. INTRODUCTION

One of the main issues of epidemiological research for the past several decades has been the relationship between biological markers (biomarkers) and disease risk. Commonly, a measurement process yields the operating characteristics of biomarkers. However, the high cost associated with evaluating these biomarkers as well as measurement errors corresponding to a measurement process can prohibit further epidemiological applications. When, for example, analysis is restricted by the high cost of assays, following Faraggi *et al.* [1], we suggest applying an efficient pooling design for collection of data. In order to allow for the instrument sensitivity problem corresponding to measurement errors, we formulate models with additive measurement errors. Obviously, these issues request assumptions on a biomarker distribution, under which operating characteristics of a biomarker can be evaluated. In this paper, we show that the distribution assumptions are related to reconstructing problems and can be formulated

in terms of characteristic functions. To this end, we assume a wide family of biospecimen distributions. The considered cases associate the measurement errors issue and evaluation based on pooled data into one stated problem.

The existence of measurement error in exposure data potentially affects any inferences regarding variability and uncertainty because the distribution representing the observed data set deviates from the distribution that represents an error-free data set. Methodologies for improving the characterization of variability and uncertainty with measurement errors in data are proposed by many biostatistical manuscripts. Thus, the model, which corresponds to observing a biomarker of interest plus a measurement error, is not in need of extensive description. However, note that, since values of a biomarker are functionally convolute with noisy measurement errors, usually distribution functions of measurement errors are assumed to be known. Moreover, on account of the complexity of extracting biomarker distribution from observed convoluted data (say deconvolution), practically, normality assumptions related to biomarker distribution are assumed. Besides, parameters of an error distribution can be evaluated by applying an auxiliary reliability study of biospecimens (e.g. a cycle of remeasuring of biospecimens: Schisterman *et al.* [2]).

Another stated problem dealing with the deconvolution is the exploration based on pooled data. Without focusing on situations where pooled data is an organic output of a study, we touch on pooling in the context of the design issue. The concept of pooling design is extensively dealt with in the statistical literature starting with publications related to cost-efficient testing of World War II recruits for syphilis. In order to reduce the cost or labor intensiveness of a study, a pooling strategy may be employed whereby 2 or more (say *p*) individual specimens are physically combined into a single 'pooled' unit for analysis. Thus, applying pooling design provides a *p*-fold decrease in the number of measured assays. Each pooled sample test result is assumed to be the average of the individual unpooled samples; this is most often the case because many tests are expressed per unit of volume (e.g. [1,3,4]). Weinberg and Umbach [5] introduced pooling as a means to estimate odds ratios for case-control studies. Although pooling design has been widely used in biological and other practice, methods for analysis of pooled data from such experiments have not been fully and well developed in the literature, except for certain special cases (e.g. Faraggi *et al.* [1] considered normal and gamma cases). This is, perhaps, partly because, for a general distribution of a biomarker, the likelihood methods based on pooled data may not be feasible since the distribution of the averages involves convolution of *p* random variables of a biomarker distribution.

The purpose of the present paper is to develop a general methodology for reasonably efficient estimation of a biomarker distribution based on data with measurement errors and/or pooled biospecimens. By utilizing the maximum likelihood technique, which is founded on characteristic functions, we consider an approach that associates the measurement error and the pooled data issues into one stated deconvolution problem. Thus, the proposed approach provides possibilities to investigate new fields related to biospecimen evaluations (e.g. a methodology for analysis of pooled data with measurement errors). For example, by applying real data, we can address issues such as whether pooling increases measurement errors, whether a physical executing of pooled data leads to errors, etc.

We believe that the demonstrated focusing on estimation of the receiver operating characteristic (ROC) curves will assist in practical applications of the methods developed in the present paper. An important step in a biomarker's development is the evaluation of its discriminating ability. The ROC curves are commonly estimated for this purpose (e.g. [6-10]). The most commonly used global index of diagnostic accuracy is the area under the ROC curve (e.g. [11]). Values of the area under the ROC curve close to 1 indicate that the marker has high diagnostic accuracy, while a value of 0.5 indicates a noninformative marker that does no better than a random (fair)

coin toss. In the case of normal-distributed data, Schisterman *et al.* [2] proposed a method to obtain the adjusted area under the ROC curve with correction for measurement error. Several authors have proposed pooling as a way to cut costs in the evaluation of biomarkers and have evaluated ROC curve analysis when dealing with such data (e.g. [1,3,4]). In the context of estimation of ROC based on data with measurement error and pooling mixtures, we analyze biomarkers from a distribution family (which includes normal distributions) that satisfy conditions related to the ability of deconvoluting (reconstructing) the target distribution.

The paper is organized as follows. Section 2 details the formal statement of the problem and a general methodology of estimation of a biomarker distribution based on convoluted data. We consider the area under the ROC curve in Section 3. To display several special aspects of the considered issue (e.g. robustness of the proposed method), Section 4 presents the results of Monte Carlo simulations. By applying the proposed technique to real pooled data, Section 5 illustrates, ROC curve evaluations. The issue of the measurement error in an actual epidemiological study is deliberated in Section 6. We exemplify the combined case regarding real pooled data with measurement errors in Section 7. Section 8 concludes the paper with several remarks.

2. FORMALIZATION AND METHOD

Let X and Y be measurements of a biomarker of interest, where X and Y are independent random samples of the disease and healthy populations, respectively. In accordance with the presented issue, practical constraints force us to weigh up the case with the observed samples in the form of (Z^X, Z^Y):

$$Z_j^X = \frac{1}{p} \sum_{i=p(j-1)+1}^{jp} X_i + J_{\varepsilon_{Xj}}, \quad Z_k^Y = \frac{1}{q} \sum_{i=q(k-1)+1}^{kq} Y_i + J_{\varepsilon_{Yk}}$$

(1)

$j = 1,..., n, k = 1,..., m$, where X_i are independent identically distributed (i.i.d.) random variables with density function f_X, Y_i are i.i.d. random variables with density function f_Y, p and q are the pool sizes in cases and controls, respectively, ε_{Xj}, ε_{Yk} are measurement errors (i.i.d. random variables, which have density functions $f_{\varepsilon X}$ and $f_{\varepsilon Y}$, respectively), and $J = 0, 1$ is a known indicator function that corresponds to the presence or absence of measurement error in the model. Mode (1) represents the mixture of notations related to the pooling design and the effect of measuring sensitivity (e.g. [2-4]). Thus, for example, the situations with $\{p = 1, q = 1, J = 1\}$, $\{p>1, q>1, J = 0\}$, and $\{p>1, q>1, J = 1\}$ coincide with the classical additive measurement error-stated problem, the pooling model definition, and the combination of pooling and additive measurement error statements, respectively (the case of $\{ p>1, q>1, J = 1\}$ can also be interpreted in terms of the presence of pooling errors). And therefore, model (1) generalizes different sample situations of biomarkers, where data can be collected in pooled sets, which are subject to additive measurement error.

The discriminating ability of biomarkers is commonly evaluated using the ROC curve, which is usually obtained by

$$\text{ROC}(t) = 1 - F_X\left(F_Y^{-1}(t)\right), \quad t \in (0, 1)$$

(2)

where F_X and F_Y are the distribution functions of X_1 and Y_1, and F^{-1} is the inverse function of F (e.g. [9]). To this end, estimation of the distribution functions of X and Y is necessary. However, estimation based on the sample (Z^X, Z^Y) instead of (X, Y) leads to the problem of

reconstructing a summand's distribution by the distribution of their sums plus the measurement error. Formally, the main focus is then to estimate f_X and f_Y based on estimators of density functions f_{Z^X} and f_{Z^Y} of random variables Z_1^X and Z_1^Y, where f_{Z^X} and f_{Z^Y} are the $p+J$- and $q+J$-folded distributions, respectively, i.e.

$$f_{Z^X}(u) = \int \cdots \int f_X\left(pu - t_1 - \cdots - t_{p-1} - pt_p\right) f_X(t_1) \ldots f_X(t_{p-1}) f_{\varepsilon X}(t_p)^J \, dt_1 \ldots dt_p$$
$$f_{Z^Y}(u) = \int \cdots \int f_Y\left(pu - t_1 - \cdots - t_{p-1} - pt_p\right) f_Y(t_1) \ldots f_Y(t_{p-1}) f_{\varepsilon Y}(t_p)^J \, dt_1 \ldots dt_p \tag{3}$$

This estimation problem is known as a *deconvolution* problem (e.g. van Es *et al.* [12]). Note that not all distribution functions can be reconstructed.

By centering in the analysis of characteristic function and deconvolution theory, Prokhorov and Ushakov [13] proposed necessary and sufficient conditions in the probability context on the distributions of *X* and *Y* such that their original distribution function can be reconstructed. Note that, even if density functions f_X and f_Y can be theoretically reconstructed, the problem of estimation of f_X and f_Y based on (Z^X, Z^Y) is very complicated. For example, if we define estimators of f_X and f_Y based on the empirical characteristic functions, then a complex definition of a *p-root* should be introduced.

A general class of the suitable distribution functions is a set of infinitely divisible distribution functions. At this rate, characteristic functions of biomarkers *X* and *Y* are uniquely determined by the characteristic functions of Z^X, Z^Y (e.g. Lukacs [14]). The infinitely divisible distributions encompass many commonly used distributions such as Normal, Cauchy, Exponential, Gamma, etc. (e.g. [14]). In the context of the parametric statement of the considered problem, given the infinitely divisible assumption, characteristic functions can be represented in the Lévy-Khinchine canonical form as

$$\phi(t) = \exp\left(it\mu + \int_{-\infty}^{\infty}\left(e^{itu} - 1 - \frac{itu}{1+u^2}\right)\frac{1+u^2}{u^2}\,dK(u)\right), \quad i = (-1)^{1/2}$$

where μ is a real unknown constant, while $K(u)$ is a nondecreasing, bounded and known (up to an unknown vector of parameters) function such that at the argument $u = -\infty$ this function is equal to zero and the integrand in $\phi(t)$ is defined for $u = 0$ by continuity to be equal to $-t^2/2$. Therefore, without loss of generality and for the sake of clarity of exposition, we assume that random variables *X* and *Y* have stable distributions with characteristic functions

$$\phi_X(t) = \phi(t; \alpha_X, \beta_X, \gamma_X, a_X), \quad \phi_Y(t) = \phi(t; \alpha_Y, \beta_Y, \gamma_Y, a_Y)$$
$$\phi(t; \alpha, \beta, \gamma, a) = e^{iat - \gamma|t|^\alpha(1 + i\beta t\omega(t,\alpha)/|t|)}, \quad t \in (-\infty, \infty)$$
$$\omega(t, \alpha) = \begin{cases} \tan(\pi\alpha/2), & \alpha \neq 1 \\ 2\ln|t|/\pi & \alpha = 1 \end{cases}$$
$$\alpha \in (0, 2], \quad |\beta| \leq 1, \quad \gamma \geq 0, \quad a \in (-\infty, \infty) \tag{4}$$

where the unknown parameters α, β, γ, and a are the characteristic exponent, a measure of skewness, the scale parameter, and the location parameter, respectively.

Definition (4) is quite general and covers many practical distributions (e.g. [15-17]), for example, the normal, $\alpha = 2$; the Cauchy, $\alpha = 1$, and the stable law of characteristic exponent, $\alpha = \frac{1}{2}$ (i.e. changing of values of the parameters can provide a modification of the classical type

of a density function corresponding to (4)). Although we consider the parametric approach to describe the characteristic functions, estimation of the unknown parameters by using directly density functions corresponding to the characteristic functions (4) is difficult; it is complicated by the fact that their densities are not generally available in closed forms (e.g. [14, p. 106]), making it difficult to apply conventional estimation methods (e.g. [17, p. 188]). Moreover, the stated issue leads to the necessity of finding the estimators based on the pooled data, and therefore, under the deconvolution problem, applying classical parametric methods is a strongly complex problem.

Since the Fourier inverse approach is a widely accepted method for the analysis of convolution and deconvolution of probability functions (e.g. [18-23]), we utilize the following well-known proposition (e.g. [17]).

Proposition 2.1

Let $\phi_\zeta(t) = E \exp(it\zeta)$ be the characteristic function of a random variable ζ, absolutely integrable. Then the corresponding distribution function F_ζ is absolutely continuous, its density function f_ζ is bounded and continuous, and

$$f_\zeta(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \phi_\zeta(t)\, dt$$

And hence by applying Proposition 2.1, we have the log-likelihood function based on the observed data

$$L_Z(\alpha,\beta,\gamma,a) = \sum_{j=1}^{n_Z} \ln\left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \exp\left(it\left(a - Z_j\right) - (s_Z)^{1-\alpha}\gamma|t|^\alpha\left(1 + i\beta t\omega\left(\frac{t}{s_Z},\alpha\right)/|t|\right)\right)\phi_{\varepsilon_Z}(t)^J dt\right)$$

(5)

where $Z = Z^X, Z^Y$; $n_Z^X = n$, $n_Z^Y = m$; $s_Z^X = p$, $s_Z^Y = q$; and $\phi_{\varepsilon_Z}^X = \phi_{\varepsilon X}$, $\phi_{\varepsilon_Z}^Y = \phi_{\varepsilon Z}^Y$ are the characteristic functions of ε_X, ε_Y, respectively. Assuming that the distributions of ε_X, ε_Y are known (perhaps, estimated by applying additional samples [2]), the maximum likelihood estimators are

$$\left\{\widehat{\alpha}_X, \widehat{\beta}_X, \widehat{\gamma}_X, \widehat{a}_X\right\} = \arg\sup_{\{\alpha_X > \varepsilon, \beta_X, \gamma_X, a_X\}} L_{Z^X}(\alpha_X, \beta_X, \gamma_X, a_X)$$
$$\left\{\widehat{\alpha}_Y, \widehat{\beta}_Y, \widehat{\gamma}_Y, \widehat{a}_Y\right\} = \arg\sup_{\{\alpha_Y > \varepsilon, \beta_Y, \gamma_Y, a_Y\}} L_{Z^Y}(\alpha_Y, \beta_Y, \gamma_Y, a_Y)$$

(6)

where ε is arbitrary small and positive. Then, as is usual for maximum likelihood estimators (e.g. [24]), estimators (6) are consistent and asymptotically normal as $n \to \infty$ and $m \to \infty$ (see for details, DuMouchel [15, p. 952]). The asymptotic variances of $\{\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{a}\}$, which are the inverses of the Fisher information matrices, are presented in the next section.

By applying (6) to Proposition 2.1, we obtain the estimators of the density functions of the biomarkers in the form

$$\widehat{f}_X(r) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it(\widehat{a}_X - r) - \widehat{\gamma}_X |t|^{\widehat{\alpha}_X}\left(1 + i\widehat{\beta}_X t\omega(t,\widehat{\alpha}_X)/|t|\right)} dt$$
$$\widehat{f}_Y(r) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it(\widehat{a}_Y - r) - \widehat{\gamma}_Y |t|^{\widehat{\alpha}_Y}\left(1 + i\widehat{\beta}_Y t\omega(t,\widehat{\alpha}_Y)/|t|\right)} dt$$

(7)

Moreover, if additional conditions $E \ln(1+|X_1|)<\infty$, $E \ln(1+|Y_1|)<\infty$ hold, *via* Theorem 1.2.4 of Ushakov [17, p. 5], we have the estimated distribution functions

$$\widehat{F}_X(u) = \frac{1}{2} - \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\mathrm{Im}\left(e^{-itu}\phi\left(t;\widehat{\alpha}_X,\widehat{\beta}_X,\widehat{\gamma}_X,\widehat{a}_X\right)\right)}{t} dt$$
$$\widehat{F}_Y(u) = \frac{1}{2} - \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{\mathrm{Im}\left(e^{-itu}\phi\left(t;\widehat{\alpha}_Y,\widehat{\beta}_Y,\widehat{\gamma}_Y,\widehat{a}_Y\right)\right)}{t} dt$$

$$(8)$$

where Im denotes the imaginary part of a function. Estimators (7) or (8) directly provide an estimator

$$\widehat{\mathrm{ROC}}(t) = 1 - \widehat{F}_X\left(\widehat{F}_Y^{-1}(t)\right), \quad t \in (0,1)$$

$$(9)$$

of the ROC curve (2).

Remark

Paulson *et al.* [16], Heathcote [25], Koutrouvelis [26], and Ushakov [17] proposed estimations of functions parameters of the stable state laws characteristic founded on the method of moments and projection methods (the integrated squares error estimator). This idea was applied to testing problems, for example, by Koutrouvelis and Kellermeier [27]. These methods of estimation are also applicable at our rate. Since pooling sizes p and q are finite, properties of these estimators (see Ushakov [17, pp. 188-197]) are preserved in the considered statement of the problem.

Estimation procedures similar to the ones proposed in this paper have been investigated by Feuerverger and McDunnough [28] in a different context.

3. ESTIMATION OF THE AREA UNDER THE ROC CURVE

The area under the ROC curve is the most commonly used summary measure of diagnostic effectiveness ([6-9,11,29], etc.). In fact Bamber [11] showed that the area under the ROC curve is equal to $P\{X>Y\}$, with values close to 1, indicating high diagnostic effectiveness. In the case of normal and gamma-distributed data, parametric approaches to estimation of the area based on the observed partial sums have been discussed by several authors (e.g. [1] as well as [3]). Since we observe samples not from target densities f_X and f_Y but from their multiple convolutions, obtaining an estimator of the area based on classical parametric methods is a complex problem. For this reason, most parametric approaches have focused on normal or gamma densities. Note that a solution of this problem also requires that the densities f_X, f_Y and $f_Z{}^X$, $f_Z{}^Y$ have one-to-one mapping.

We consider two ways of representing the area. Define $A \equiv P\{X>Y\}$ and then directly we have

$$A = \int_{-\infty}^{\infty} P\{Y<u\} f_X(u)\, du$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{u} f_Y(r)\, dr\, f_X(u)\, du$$

$$(10)$$

However, if the expectation $E \ln(1 + |X_1 - Y_1|)<\infty$, by Theorem 1.2.4 of Ushakov [17, p. 5], we obtain a simpler expression

$$A = \frac{1}{2} + \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{\phi_X(t)\,\phi_Y(-t) - \phi_X(-t)\,\phi_Y(t)}{t} dt$$

$$(11)$$

Thus, combining (10) with (7) or (11) with (4), (6) yields the estimator $\hat{A}$. Since the maximum likelihood estimators of the unknown parameters are utilized, we can evaluate the asymptotic distribution of $\hat{A}$ by applying the usual Taylor expansion. Following Kotz *et al.* [30, p. 122], we conclude that if the sample sizes *n, m* satisfy $0< \lim_{n+m \to \infty} m/(n + m) = \rho<1$, then

$$(n+m)^{1/2} \left(\widehat{A} - A\right) \mathrm{N}\left(0, \sigma_A^2\right) \quad \text{as} \quad n, m \to \infty$$

(12)

where the asymptotic variance σ_A^2 is

$$\sigma_A^2 = \frac{1}{1-\rho} \sum_{i=1}^{4} \sum_{j=1}^{4} W_{ij}^{(X)} \frac{\partial A}{\partial \theta_i^{(X)}} \frac{\partial A}{\partial \theta_j^{(X)}} + \frac{1}{\rho} \sum_{i=1}^{4} \sum_{j=1}^{4} W_{ij}^{(Y)} \frac{\partial A}{\partial \theta_i^{(Y)}} \frac{\partial A}{\partial \theta_j^{(Y)}}$$
$$\theta^{(X)} = [\alpha_X, \beta_X, \gamma_X, a_X], \quad \theta^{(Y)} = [\alpha_Y, \beta_Y, \gamma_Y, a_Y]$$

where $\left\{W_{ij}^{(X)}\right\}^{-1}$ and $\left\{W_{ij}^{(Y)}\right\}^{-1}$ are the Fisher information matrices of the estimators $\{\hat{\alpha}_X, \hat{\beta}_X, \hat{\gamma}_X, a\hat{\ }_X\}$ and $\{\hat{\alpha}_Y, \hat{\beta}_Y, \hat{\gamma}_Y, a\hat{\ }_Y\}$, respectively, i.e.

$$\left\{W_{ij}^{(X)}\right\}^{-1} = \left\{ -E \frac{\partial^2 \ln \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itZ^X} \phi_X\left(t/p; \theta^{(X)}\right)^p \phi_{\varepsilon_X}(t)^J dt}{\partial \theta_i^{(X)} \partial \theta_j^{(X)}} \right\}$$

$$\left\{W_{ij}^{(Y)}\right\}^{-1} = \left\{ -E \frac{\partial^2 \ln \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itZ^Y} \phi_Y\left(t/p; \theta^{(Y)}\right)^p \phi_{\varepsilon_Y}(t)^J dt}{\partial \theta_i^{(Y)} \partial \theta_j^{(Y)}} \right\}$$

(In the case of $J = 0$, Fisher information matrices $\left\{W_{ij}^{(X)}\right\}^{-1}$ and $\left\{W_{ij}^{(Y)}\right\}^{-1}$ are fully investigated by DuMouchel [31].)

4. MONTE CARLO SIMULATION STUDY

4.1. Normally distributed data

To apply the proposed method, the assumption of normal distribution is not necessary. Obviously, in the case where $J = 0$ in (1) and measurements of a biomarker are known to be normally distributed, the technique presented by Faraggi *et al.* [1] provides a more efficient estimation of an ROC than the approach based on stable distribution. Here, we perform the following Monte Carlo simulations. Let us assume that one believes the observations are normally distributed and chooses the method of [1]. Alternatively, we apply ROC curve estimation as proposed in Section 2. However, the true diagnostic markers satisfy (1), where

$J = 0$, $X_i \tilde{} N(0.8, 1)$, $Y_i \tilde{} N(0, 1)$, and $i = 1,..., N$. We ran 2000 repetitions of the sample $\{Z_j^X, Z_k^Y, 1 \le j, k \le n\}$ at each argument $u = -3, -2.55,..., 2.44, 2.89, 3.356$ of the distribution functions F_X, F_Y, and pooling group size $p = q = 2, 4, 5$ (where $n = N/p$). Denote the ratios

$$\Delta_Y(u) = \frac{\overline{\left(\widehat{F}_Y(u) - \Phi(u)\right)^2}}{\overline{\left(\tilde{F}_Y(u) - \Phi(u)\right)^2}} \quad \text{and} \quad \Delta_X(u) = \frac{\overline{\left(\widehat{F}_X(u) - \Phi(u - 0.8)\right)^2}}{\overline{\left(\tilde{F}_X(u) - \Phi(u - 0.8)\right)^2}}$$

where $\bar{a}$ is the Monte Carlo expectation of a; estimators $\hat{F}_X$, $\hat{F}_Y$ and $\tilde{F}_X$, $\tilde{F}_Y$ are defined by (8) and [1], respectively; and Π is the standard normal distribution function. In this case, Table I presents the relative efficiency of the proposed estimation in terms of Δ_X and Δ_Y.

Thus, when the information regarding the biomarker distribution is ignored (or called into doubt), the proposed method provides less efficient estimation of the tails belonging to the distributions than that based on Faraggi *et al.*'s [1] approach. In accordance with Table I, this loss in efficiency asymptotically decreases as $N \to \infty$.

In this simulation study, the AUC is equal to $A = P\{X > Y\} \simeq 0.714$. Following (12), the variance of the estimator $\hat{A}$ can be approximated by $\sigma_A^2 / (2N/p) \simeq 0.1384 / (2N/p)$. In the cases of ($N = 300$, $p = 2$), ($N = 200$, $p = 2$), and ($N = 100$, $p = 2$), σ_A^2's were estimated by substituting of $\theta^{(X)}$ and $\theta(Y)$ by $[\hat{\alpha}_X, \hat{\beta}_X, \hat{\gamma}_X, a\char`\^_X]$ and $[\hat{\alpha}_Y, \hat{\beta}_Y, \hat{\gamma}_Y, a\char`\^_Y]$, and the Monte Carlo means $\overline{\sigma_A^2}$ (and variances $\overline{2N/p\left(\hat{A}-A\right)^2}$) were 0.1383 (0.1387), 0.1431 (0.1472), and 0.1474 (0.1685), respectively. (By Table 1 of [1], we can approximate

$$2N/p\overline{\left(\hat{A}-A\right)^2} \simeq (2 \times 200/2)\,(0.026)^2 = 0.1352 \text{ and}$$

$$2N/p\overline{\left(\hat{A}-A\right)^2} \simeq (2 \times 100)\,/2(0.037)^2 = 0.1369 \text{ when the unknown parameters are estimated}$$

under executed normal distributional assumptions.) Thus, when pooled data have less than 50 observations (Z^X, Z^Y), application of the asymptotic result (12) can lead to biased estimations of $\tilde{A}$'s variance.

Note that, when the information regarding normal distributions of X and Y in (1) is accepted, the special form of the characteristic function (4) leads to the estimation scheme that coincides with the method of Faraggi *et al.* [1].

4.2. Robustness

The simulations thus far assumed that the samples followed normal distributions. In order to illustrate the robustness of our methodology, in a similar manner to the previous paragraph, we perform the Monte Carlo simulations based on X's and Y's from t-distributions with degrees of freedom (df) and means 0 and 0.8, respectively. Figure 1 depicts graphs

$$\overline{\left(1 - \hat{F}_Y(u),\ 1 - \hat{F}_X(u)\right)} \text{ and } \overline{\left(1 - \tilde{F}_Y(u),\ 1 - \tilde{F}_X(u)\right)} \text{ that are the Monte Carlo expectations of}$$

the estimators of the ROC curve ($1 - F_Y(u)$, $1 - F_X(u)$), which $u = -3, -2.55, \ldots, 2.44, 2.89, 3.356$.

In comparing with Table I, for example, when $p = 2$ and df $= 5$, $\Delta_X(u)$ and $\Delta_Y(u)$ are close to 0.989 for $u \leq -2$ and $u \geq 2.4$, whereas $0.351 \leq \Delta_X(u)$, $\Delta_Y(u) \leq 0.900$ for $u \in (-2, 2.4)$. Thus, we can conclude that the proposed method is more robust than the approach based on normal distributional assumptions.

5. POOLING STUDY

We exemplify the proposed methodology for the pooling design (i.e. $J = 0$ in (1)) from a case-control study of a biomarker of coronary heart disease. The *interleukin*-6 biomarker of inflammation has been suggested as having potential discriminatory ability for myocardial infarction (MI). However, since the cost of a single assay is 74 dollars, examination of its usefulness has been hindered. Hence, it is reasonable to investigate the effectiveness of pooling. In this study, the inflammation marker *interleukin*-6 (measured in mg/dl) was obtained from

$n = 40$ cases who recently survived an MI and from $m = 40$ controls who had a normal rest electro cardiogram (ECG), were free of symptoms and had no previous cardiovascular procedures or MIs. In order to evaluate the success of the cost-efficient pooling design, blood specimens were randomly pooled in groups of $p = q = 2$ for the cases and the controls separately, and *interleukin*-6 was re-measured and treated as the average of the corresponding individual *interleukin*-6. The objective of the study is to evaluate, in the context of ROC curve analysis, the discriminating abilities of the biomarker in identifying patients at high risk of coronary heart disease. Throughout the section, *X* and *Y* represent, respectively, individual *interleukin*-6 measurements from a diseased and a healthy subject.

5.1. The unpooled (full) data

Preliminarily note that, empirically, the mean and standard deviation of *X* and *Y* are ($\bar{X} = 4.288$, $\bar{\sigma}_X = 2.176$) and ($\bar{Y} = 1.846$, $\bar{\sigma}_Y = 1.366$), respectively. Consider the problem of finding an ROC estimator based on the full data in the context of Section 2, where $p = q = 1$, $J = 0$, and the characteristic functions are represented in terms of (4). "R: A Programming Environment for Data Analysis and Graphics" is utilized, in order to solve the optimization problem (6), where we define the initial parameters for ($\hat{\alpha}_X$, $\hat{\beta}_X$, $\hat{\gamma}_X$, $\hat{a}_X$) and ($\hat{\alpha}_Y$, $\hat{\beta}_Y$, $\hat{\gamma}_Y$, $a\,\hat{}_Y$) as (2, 0, $\bar{\sigma}_x^2/2$, $\bar{X}$) and (2, 0, $\bar{\sigma}_y^2/2$, $\bar{Y}$). (Schematic R codes are mentioned in the Appendix A.) As a result, we have ($\hat{\alpha}_X$, $\hat{\beta}_X$, $\hat{\gamma}_X$, $a\,\hat{}_X$) = (1.990, -0.414, 2.309, 4.288) and ($\hat{\alpha}_Y$, $\hat{\beta}_Y$, $\hat{\gamma}_Y$, $a\,\hat{}_Y$) = (1.405, -0.901, 0.629, 2.201). Hence, the estimated density of *X* is very close to normal and the nonparametric test results presented in Table II coincide with this fact.

The estimation procedure (6), (8) leads to the estimators of the distribution functions of *X* and *Y*, which are plotted in graphs (a) and (b) of Figure 2.

On the basis of (9), Figure 2(c) depicts the estimated ROC curve based on the empirical distribution functions and the proposed estimators based on unpooled data.

According to the literature in Section 3, we calculate the area under the ROC curve using the Wilcoxon *U* statistic as about 0.831 ± 0.046. By applying the proposed methodology, the area is estimated to be 0.818. Therefore, biomarker *interleukin*-6 has strong discriminating ability in identifying patients at high risk of coronary heart disease. At the same time, at this rate, the proposed methodology is reasonable but not so necessary because, for example, nonparametric techniques can be utilized. The results presented here will be compared with the following outputs.

5.2. The pooled data

Empirical characteristics based on the pooled data are

$\bar{X} = \bar{Z}^X = 4.342, \widehat{\sigma}_x = p^{1/2}\,\mathrm{SD}\left(Z^X\right) = 2.266$ and $\bar{Y} = \bar{Z}^Y = 1.701, \widehat{\sigma}_y = q^{1/2}\,\mathrm{SD}\left(Z^Y\right) = 1.234$. In a similar manner, with $p = q = 2$, $J = 0$ in (1), we estimate (based on the pooled sample) the parameters as ($\hat{\alpha}_X$, $\hat{\beta}_X$, $\hat{\gamma}_X$, $a\,\hat{}_X$) = (1.997, -0.038, 2.439, 4.342) and ($\hat{\alpha}_Y$, $\hat{\beta}_Y$, $\hat{\gamma}_Y$, $a\,\hat{}_Y$) = (1.963, -0.410, 0.723, 1.701). Now, the estimated densities of *X* and *Y* are both approximately normal density functions. For example, Table II presents nonparametric tests for normality of Z^Y. The estimation procedure (8) yields the estimators of the distribution functions of *X* and *Y*, which are plotted in Figure 2. Accordingly, Figure 2 also depicts the estimated ROC curve based on the empirical distribution functions and the proposed estimation based on the pooled data. By (11), we obtain $\hat{A} = 0.841$. In this way, we demonstrate that use of the proposed methodology and the cost-efficient pooling design provides statistical conclusions similar to the results based on full data.

6. STUDY OF MEASUREMENT ERROR

In the context of a practical study, Schisterman *et al.* [2] proposed model (1) with normally distributed X and Y, assuming $p = q = 1$, $J = 1$. We briefly describe the study as follows. A thiobarbituric acid reaction substance (TBARS) is a biomarker that measures sub-products of lipid peroxidation and has been proposed as a discriminating measurement between cardiovascular disease cases (X) and healthy controls (Y). Blood samples, physical measurements, and a detailed questionnaire on different behavioral and physiological patterns were obtained from study participants. The cases are defined as individuals with MI. Due to skewness of the original data, the transformation $(TBARS)^{-2}$ was implemented in order to bring the data distribution closer to normality. Empirical evaluations based on the transformed data are $n = 47$, $\bar{X} = 0.411$, $SD(X) = 0.260$, and $m = 891$, $\bar{Y} = 0.608$, $SD(Y) = 0.302$. In order to analyze the influence of measurement errors (ε_X, ε_Y), additional data were obtained by an extra sampling. A reliability study was conducted on a convenience sample of 10 participants. Twelve-hour fasting blood samples were obtained in seven women and three men over a period of six months. The blood samples were obtained every month on the same day of each female's menstrual cycle and every month on the same calendar day for each male. Under the natural assumption that the distribution functions of ε_X, ε_Y are equal and normal, from the reliability study we obtain that ε_X has zero mean and 0.0567 variance. Without correction for measurement error, an estimator of A (the area under the ROC curve) based on Z^X and Z^Y is 0.699, with the estimated 95 per cent confidence interval being (0.616, 0.782). After correcting for the measurement error, Schisterman *et al.* [2] obtained the adjusted area estimate to be 0.735, with the 95 per cent confidence interval being (0.582, 0.888). Noting that the situation in which X and Y have normal distributions is the special case of the proposed methodology, in a similar manner to Section 5, we utilize the technique from Sections 2, and 3. We have $(\hat{\alpha}_X, \hat{\beta}_X, \hat{\gamma}_X, \hat{a}_X) = (1.349, -1.002, 0.013, 0.430)$ and $(\hat{\alpha}_Y, \hat{\beta}_Y, \hat{\gamma}_Y, \hat{a}_Y) = (1.888, -0.103, 0.010, 0.597)$. By (9), Figure 3 graphically displays these results, which lead to $\hat{A} = 0.831$, with the estimated asymptotical 95 per cent confidence interval being (0.664, 0.998).

Therefore, the correction for measurement error and the use of the proposed method increased the estimator of the area under the ROC curve and shifted the confidence interval to include much higher values. Use of the uncorrected results underestimates the effectiveness of TBARS as a biomarker capable of discriminating between subjects with and without cardiovascular disease.

7. MIXED STUDY

Let us consider the study that is described in Section 5 with $J = 1$ in model (1). Assume that measurement errors ε_X, ε_Y have identical distribution functions. Therefore, the fulfillment of the sampling strategy of the study yields observations

$$Z_j^X = \frac{X_{2(j-1)+1} + X_{2j}}{2} + \varepsilon_j, \quad Z_k^Y = \frac{Y_{2(k-1)+1} + Y_{2k}}{2} + \varepsilon_k', \quad j = 1, \dots, 20, \quad k = 1, \dots, 20$$
$$Z_j'^X = X_j + \varepsilon_j'', \quad Z_k'^Y = Y_k + \varepsilon_k''', \quad j = 1, \dots, 40, \quad k = 1, \dots, 40$$

where ε, ε', ε'', and ε''' are i.i.d. random variables. It is natural to assume that ε, ε', ε'', and ε''' have characteristic function $\phi(t; \alpha, \beta, \gamma, a)$, which is defined by (4) with $a = 0$ (e.g. a distribution function of ε is a normal with zero mean). Hence, the characteristic function of the observed i.i.d. random variables

$$T_j = \begin{cases} Z_j^X - \dfrac{Z'^X_{2(j-1)+1} + Z'^X_{2j}}{2} & \left(= \varepsilon_j - \dfrac{\varepsilon''_{2(j-1)+1} + \varepsilon''_{2j}}{2}\right), & j \leq 20 \\[3mm] Z_{j-20}^Y - \dfrac{Z'^Y_{2(j-21)+1} + Z'^X_{2(j-20)}}{2} & \left(= \varepsilon_j - \dfrac{\varepsilon'''_{2(j-1)+1} + \varepsilon'''_{2j}}{2}\right), & 20 < j \leq 40 \end{cases}$$

is $\phi_T(t; \alpha, \beta, \gamma) = \exp(-\gamma|t|^\alpha (1 + i\beta t\omega(t, \alpha)/|t|) - 2\gamma|t/2|^\alpha (1 - i\beta t\omega(-t/2, \alpha)/|t|))$. By virtue of the proposed method (here,

$$L_T(\alpha, \beta, \gamma) = \sum_{j=1}^{40} \ln\left(\frac{1}{2\pi} \int_{-\infty}^{\infty} \exp\left(-itT_j\right) \phi_T(t; \alpha, \beta, \gamma)\, dt\right)$$

is shown to be the log likelihood), we estimate the unknown parameters to be $\hat{\alpha} = 1.925$, $\hat{\beta} = -0.981$, $\hat{\gamma} = 0.162$ and represent the estimated distribution function of the measurement errors in Figure 4.

Since T and ε are not identically distributed, in this study, the issue of reconstructing the target distribution, which is considered in Section 2, is also in effect. In this practical example, we have considered the pooled data in the context of resampling to evaluate the problem related to instrument sensitivity. Note that applying the obtained result as in Section 6 to the correction of the ROC estimation does not provide palpable outcomes ($\hat{A} = 0.831$ is corrected up to 0.842). However, the example of alternative use of pooling design makes sense in the aspect of detection of the measurement error.

8. DISCUSSION

The paper demonstrates that the deconvolution issue is a very important field of practical and theoretical biostatistics. Certainly, considerations of non-normal distributions by way of reconstructed solutions of the deconvolution problem are reasonable topics. However, it is necessary to allow for the fact that the class of appropriated distribution functions is bounded by theoretical conditions related to the ability of reconstructing a target distribution. Due to this reason, practically, we can consider the family of infinitely divisible distribution functions.

In the present manuscript, we apply the stable distribution functions, which are a valuable subclass of infinitely divisible distributions. Even in the case of stable distributions, the proposed maximum likelihood method extends the classical parametric approach, because, in a sense, the classical type of the target distribution function is also estimated, i.e. for example, if the α estimated by (6) is close to 2 or 1, we can conclude that the estimated density is approximately normal or Cauchy, respectively. Besides, the considered density functions are not generally available in closed analytical forms.

Modern statistical software (e.g. R, S-PLUS) provides simplicity in applying the proposed methodology to the real data studies.

ACKNOWLEDGEMENTS

APPENDIX A: SCHEMATIC R CODES

Without loss of generality and for the sake of clarity of exposition, we assume $J = 0$ and that p is the pool size in (1). The characteristic function (4) of observations Z^X (or Z^Y) can be coded in the form

```
fi<-function(t,alpha,beta,gamma,a){

    if (alpha==1) w<-2*log(abs(t)/p)/pi else w<-tan(pi*alpha/2)

    return(exp(1i*a*t-(p^(1-alpha)*gamma*abs(t)^(alpha))*(1+1i*beta*sign(t)*w))) }
```

where (alpha,beta,gamma,a) are parameters of the characteristic function of X (or Y)and p is a known fixed parameter corresponding to the pool size. Following Proposition 2.1, we have the density function

```
fs<-function(u,alpha,beta,gamma,a){

    integ1<-function(t) exp(-1i*u*t)*fi(t,alpha,beta,gamma,a)

    integ1R<-function(t) Re(integ1(t))

    integrate(integ1R,-Inf,Inf)[[1]]/(2*pi) }
```

Assume that L(alpha, beta, gamma, a) is the log likelihood function (5) based on fs at sample Z. Denote LV<-function(r) -L(r[1],r[2],r[3],r[4])and estimate (alpha, beta,gamma,a) at Estim \$par, where

```
Estim<-optim(par=c(alpha0,beta0,gamma0,a0) ,LV,lower =c(alphal,betal,gammal,al), upper
= c(alphau,betau,gammau,au))
```

par, lower and upper are initial, lower and upper bound values for the parameters, respectively.

REFERENCES

1. Faraggi D, Reiser B, Schisterman EF. ROC curve analysis for biomarkers based on pooled assessments. Statistics in Medicine 2003;22:2515–2527. [PubMed: 12872306]

2. Schisterman EF, Faraggi D, Reiser B, Trevisan M. Statistical inference for the area under the receiver operating characteristic curve in the presence of random measurement error. American Journal of Epidemiology 2001;154:174–179. [PubMed: 11447052]

3. Liu A, Schisterman EF. Comparison of diagnostic accuracy of biomarkers with pooled assessments. Biometrical Journal 2003;45:631–644.

4. Liu A, Schisterman EF, Theo E. Sample size and power calculation in comparing diagnostic accuracy of biomarkers with pooled assessments. Journal of Applied Statistics 2004;31:49–59.

5. Weinberg CR, Umbach DM. Using pooled exposure assessment to improve efficiency in case-control studies. Biometrics 1999;55:718–726. [PubMed: 11314998]

6. Wieand S, Gail MH, James BR, James KL. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired Data. Biometrika 1989;76:585–592.

7. Goddard MJ, Hinbery I. Receiver operator characteristic (ROC) curves and non-normal data: an empirical study. Statistics in Medicine 1990;9:325–337. [PubMed: 2343225]

8. Zweig MH, Campbell G. Receiver operator characteristic (ROC) plots; a fundamental evaluation tool in clinical medicine. Clinical Chemistry 1993;39:561–577. [PubMed: 8472349]

9. Shapiro DE. The interpretation of diagnostic tests. Statistical Methods in Medical Research 1999;8:113–134. [PubMed: 10501649]

10. Zhou, XH.; Obuchowski, NA.; McClish, DK. Statistical Methods in Diagnostic Medicine. Wiley; New York: 2002.

11. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology 1975;12:387–415.

12. van Es B, Jongbloed G, van Zuijlen M. Isotonic inverse estimators for nonparametric deconvolution. Annals of Statistics 1998;26:2395–2406.

13. Prokhorov AV, Ushakov NG. On the problem of reconstructing a summands distribution by the distribution of their sum. Theory of Probability and its Applications 2002;46:420–430.

14. Lukacs, E. Characteristic Functions. Griffin; London: 1970.

15. DuMouchel WH. On the asymptotic normality of the maximum-likelihood estimate when sampling from a stable distribution. Annals of Statistics 1973;1:948–957.

16. Paulson AS, Holcomb EW, Leitch RA. The estimation of the parameters of the stable laws. Biometrika 1975;62:163–170.

17. Ushakov, NG. Modern Probability and Statistics. VSP VB; Utrecht, The Netherlands: 1999. Selected topics in characteristic functions.

18. Halász G, Major P. Reconstructing the distribution from partial sums of samples. Annals of Statistics 1977;5:987–998.

19. Fan J. On the optimal rates of convergence for nonparametric deconvolution problems. Annals of Statistics 1991;19:1257–1272.

20. Diggle PJ, Hall P. A Fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society*, *Series B* 1993;55:523–531.

21. Fotopoulos SB. Invariance principles for deconvolving kernel density estimation for stationary sequences of random variables. Journal of Statistical Planning and Inference 2000;86:31–50.

22. Koltchinskii VI. Empirical geometry of multivariate data: a deconvolution approach. Annals of Statistics 2000;28:591–629.

23. Watteel RN, Kulperger RJ. Nonparametric estimation of the canonical measure for infinitely divisible distributions. Journal of Statistical Computation and Simulation 2003;73:525–542.

24. Serfling, RJ. Approximation Theorems of Mathematical Statistics. Wiley; New York: 1980.

25. Heathcote CR. The integrated squared error estimation of parameters. Biometrika 1977;64:255–264.

26. Koutrouvelis IA. Regression-type estimation of the parameters of stable laws. Journal of the American Statistical Association 1980;75:918–928.

27. Koutrouvelis IA, Kellermeier J. A goodness-of-fit test based on the empirical characteristic function when parameters must be estimated. *Journal of the Royal Statistical Society*, *Series B* 1981;43:173–176.

28. Feuerverger A, McDunnough P. On some fourier methods for inference. Journal of the American Statistical Association 1981;76:379–387.

29. Zou KH, Hall WJ, Shapiro DE. Smooth non-parametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine 1997;16:2143–2156. [PubMed: 9330425]

30. Kotz, S.; Lumelskii, Y.; Pensky, M. *The Stress-strength Model and its Generalizations*: *Theory and Applications*. World Scientific; Singapore: 2000. p. 1-272.

31. DuMouchel WH. Stable distribution in statistical inference: 2. Information from stably distributed samples. Journal of the American Statistical Association 1975;70:386–393.
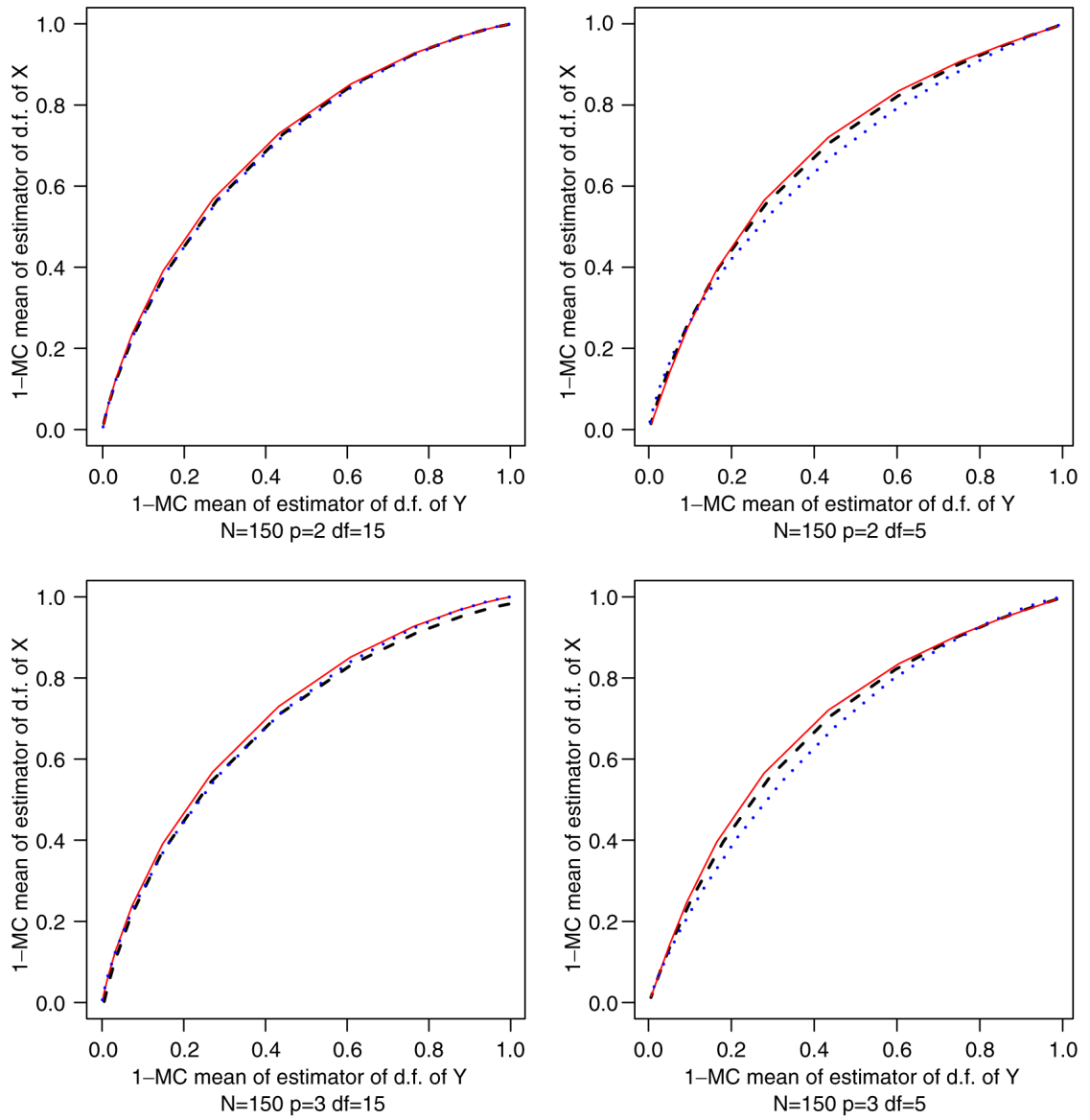
Figure 1.
Comparison between the ROC curve based on the *t*-distribution functions of *X* and *Y* (curve
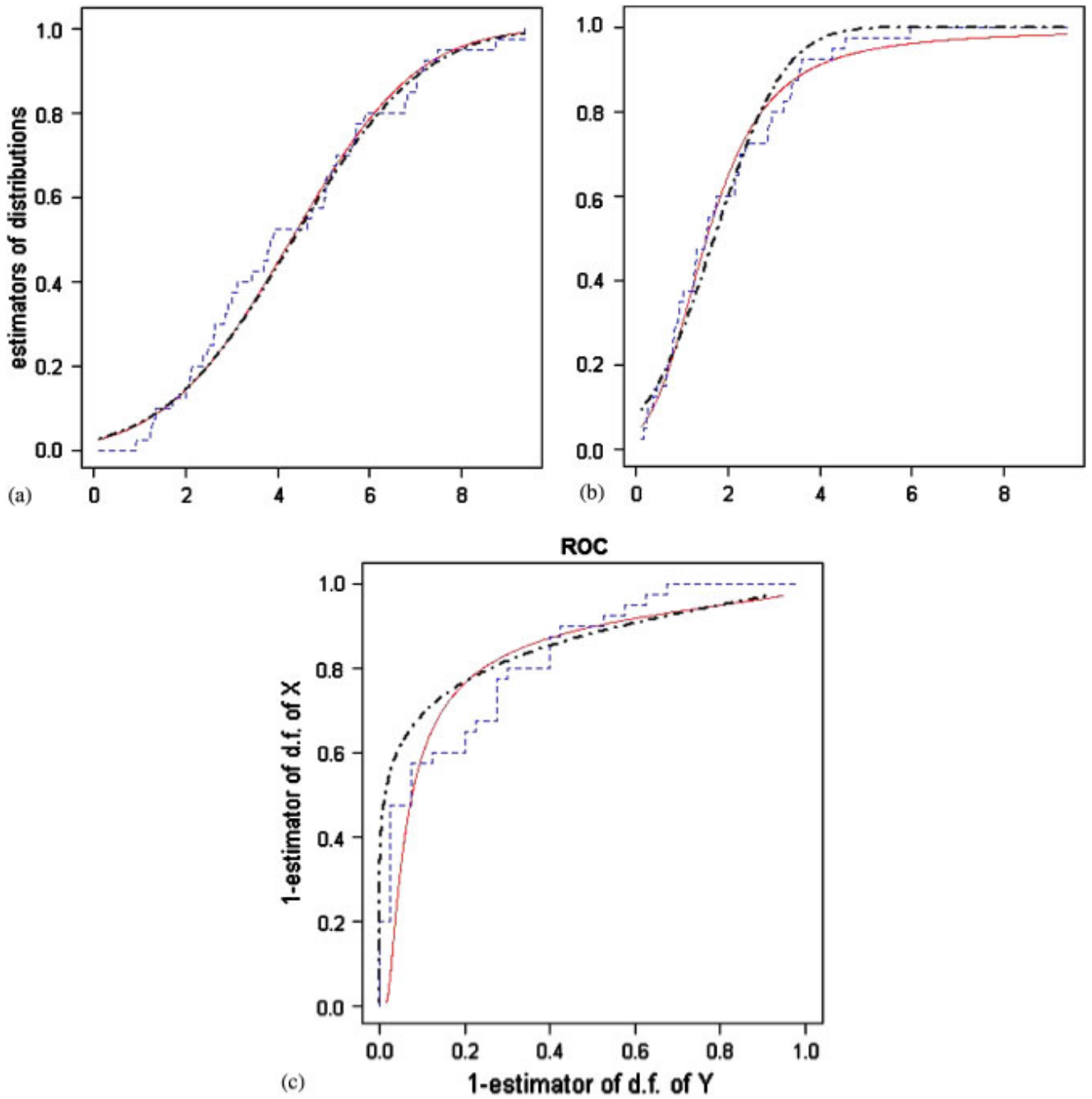—) and the ROC estimators by (9) (- - -) and Faraggi *et al.* [1] (⋯).

Figure 2.
Comparison between the empirical distribution functions of *X, Y*, the empirical ROC curve
(curves - - - - of the graphs (a)-(c), respectively) and the corresponding model-based estimators
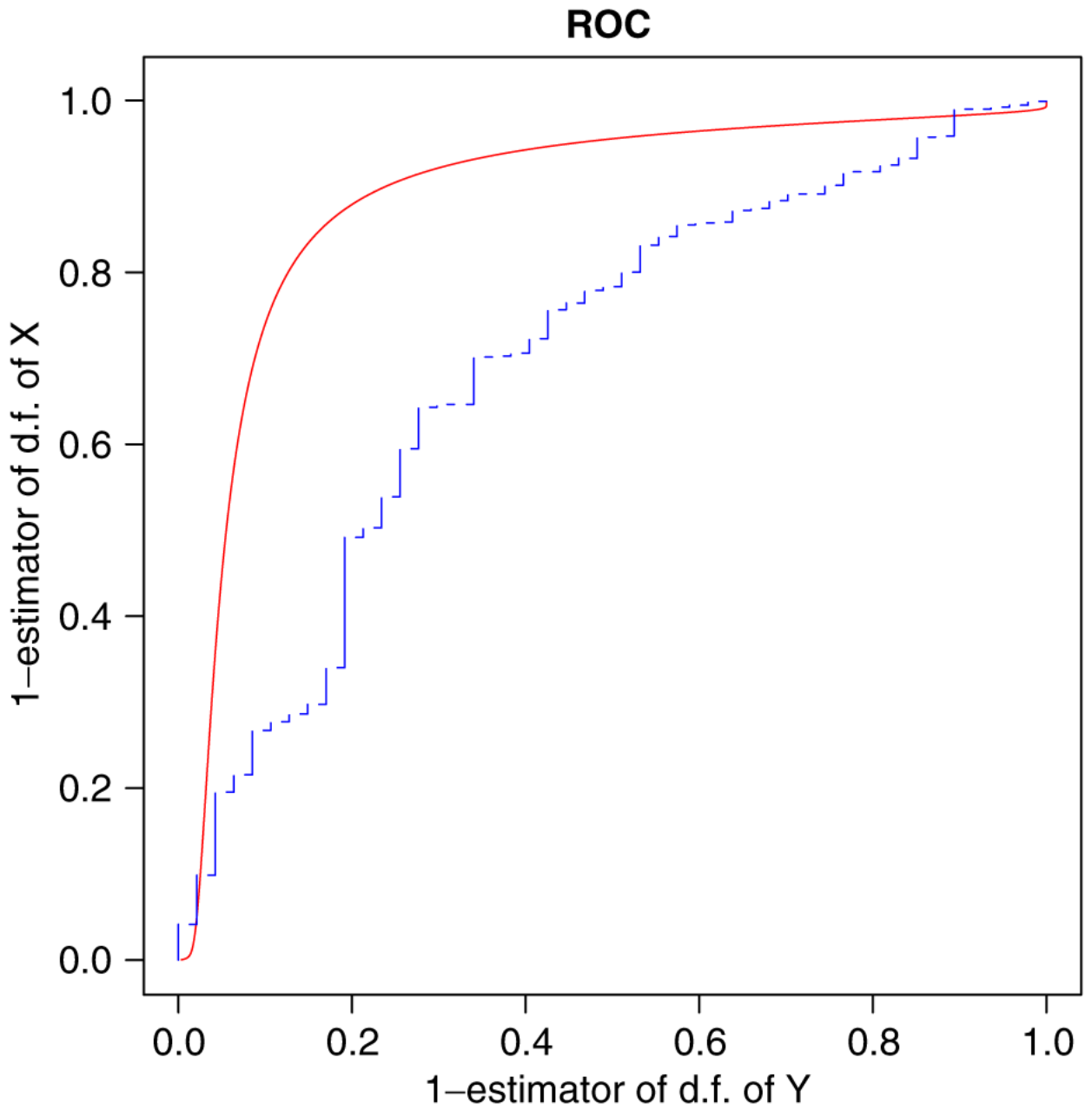(curves— and - · -), which are based on unpooled and pooled data, respectively.

Figure 3.
Comparison between the ROC curve based on the empirical distribution functions of Z^X and Z^Y (- - -) and the adjusted ROC curve estimator (—).
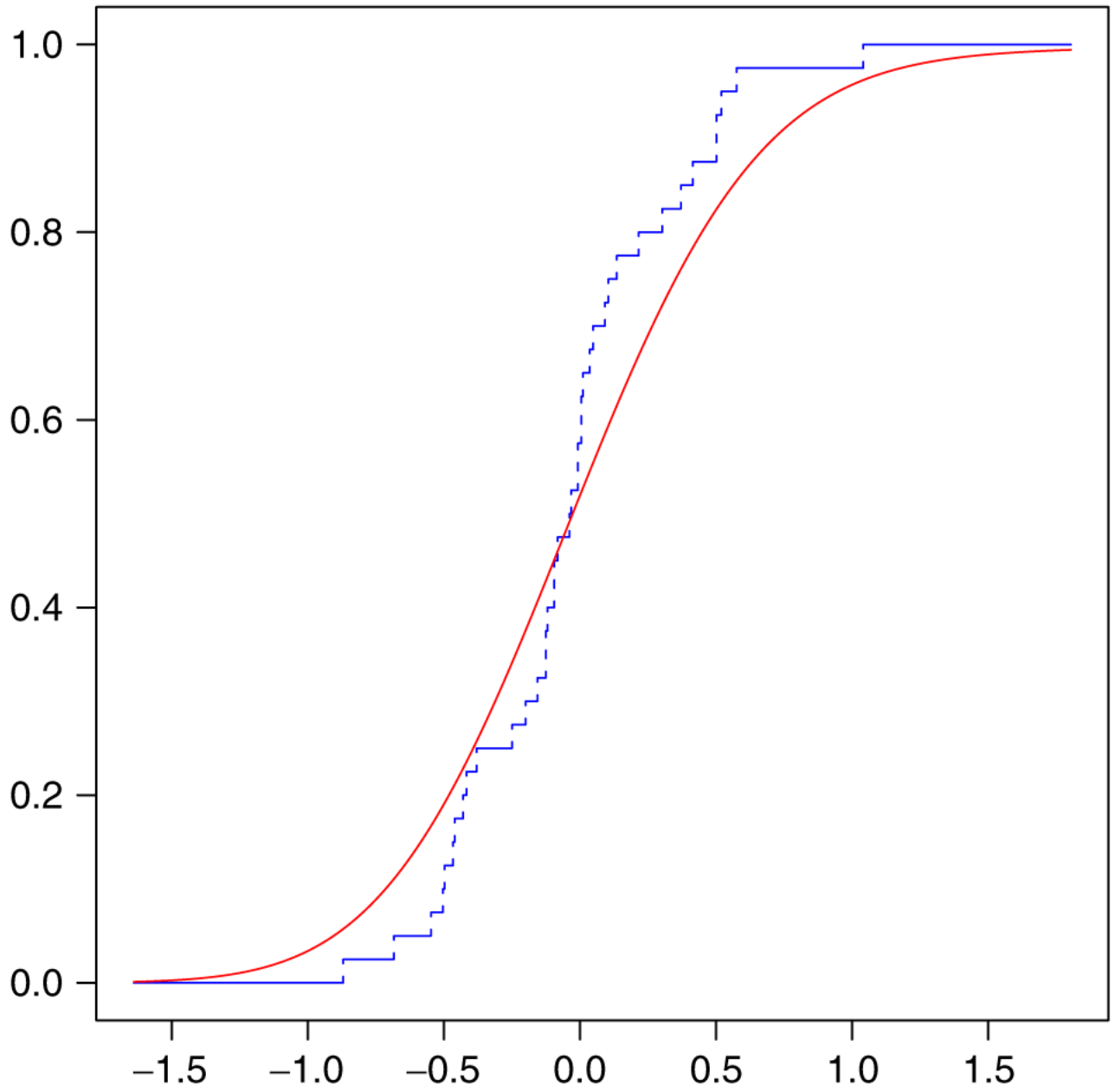
Figure 4.
Curves — and - - - present the estimated distribution function of measurement errors ε_j and the empirical distribution function of observed random variables T_j, respectively.

Table I

The ratios of the Monte Carlo variances related to the estimators of $F_X(u)$ and $F_Y(u)$

u	N = 100, p = 2		N = 100, p = 4		N = 300, p = 2	
	$\Delta_Y(u)$	$\Delta_X(u)$	$\Delta_Y(u)$	$\Delta_X(u)$	$\Delta_Y(u)$	$\Delta_X(u)$
-3	1.999	2.551	1.976	2.955	1.775	1.903
-2.547	1.638	1.733	1.065	1.153	1.591	1.394
-2.093	1.034	1.121	0.891	0.999	1.087	1.107
-1.640	0.990	1.109	0.995	0.998	0.990	1.004
-1.187	1.075	1.044	1.215	1.198	1.039	1.021
-0.733	1.166	1.206	1.407	1.511	1.162	1.251
-0.280	1.255	1.247	1.607	1.612	1.299	1.311
0.173	1.255	1.278	1.614	1.711	1.253	1.321
0.627	1.166	1.209	1.416	1.541	1.099	1.008
1.080	1.090	1.107	1.128	1.119	1.028	1.011
1.533	1.022	1.111	0.923	1.005	1.001	1.102
1.987	1.023	1.035	0.922	1.012	1.007	1.014
2.440	1.385	1.393	1.351	1.255	1.255	1.199
2.893	2.119	2.422	2.571	2.482	1.784	1.898
3.347	2.689	3.039	3.036	2.953	2.053	2.147

Table II

Tests for normality of X and Z^Y ((a) and (b), respectively)

Test statistic	(a)		(b)	
	Value	*p*-value	Value	*p*-value
Shapiro-Wilk	0.961368	0.1865	0.951419	0.3891
Kolmogorov-Smirnov	0.105298	>0.1500	0.130382	>0.1500
Cramer-von Mises	0.071906	>0.2500	0.057138	>0.2500
Anderson-Darling	0.461674	0.2489	0.373652	>0.2500