On nonparametric inference for $P(X < Y)$ for paired variables

J. A. Montoya[*] and F. J. Rubio[†]

Abstract

We propose a class of nonparametric point estimators for $\theta = P(X < Y)$ for the case where (X, Y) are paired, possibly dependent, continuous random variables. We make use of the pairing structure for linking the estimation of θ with the estimation of the survival function and density function of $Y - X$. We consider the use of bootstrap to obtain confidence intervals for θ based on the proposed estimators. The performance of these estimators is illustrated using simulated and real data. The example with real data shows that not accounting for pairing and dependence might lead to different conclusions about the relationship between X and Y.

Key Words: Bootstrap, dependence, nonparametric, paired observations, stress-strength model.

1 Introduction

The study of stress–strength models have received considerable attention for many years due to its applicability in diverse areas. The main interest in this kind of models is the quantity $\theta = P(X < Y)$, where X and Y are random variables. In medicine for example, if X and Y are the outcomes of a control and an experimental treatment respectively, the parameter θ can be interpreted as the effectiveness of treatment Y (Ventura et al., 2011). This quantity is also related to the Receiver Operating Characteristic (ROC) curves, where θ is interpreted as an index of accuracy (Zhou, 2008). In engineering and reliability studies θ is also a quantity of interest because it may represent the probability that the strength of a component (Y) exceeds the stress (X) coming from external factors (Kotz et al., 2003).

Stress-strength models were introduced by Birnbaum (1956) who proposed a nonparametric estimator for θ based on the Mann-Whitney statistic for the case where X and Y are independent. There is a large amount of literature related to the study of point and interval estimation

[*] UNIVERSIDAD DE SONORA, DEPARTAMENTO DE MATEMÁTICAS. E-mail: montoya@mat.uson.mx

[†] UNIVERSITY OF WARWICK, DEPARTMENT OF STATISTICS, COVENTRY, CV4 7AL, UK. E-mail: F.J.Rubio@warwick.ac.uk

of θ using different approaches (see Kotz et al., 2003 for a good survey on this). For instance, in the case where X and Y are independent, Sun et al. (1998) proposes a Bayesian approach using reference priors; Baklizi and Eidous (2006) propose an estimator based on kernel estimators of the densities of X and Y (which can be straightforwardly generalised to the use of other nonparametric density estimators); Zhou (2008) proposes the use of bootstrap and asymptotic intervals; Jing et al. (2009) estimate θ using the empirical likelihood; Montoya (2008) and Díaz–Francés and Montoya (2012) propose the use of the profile likelihood for conducting inference about θ; and Ventura et al. (2011) propose the use of Bayesian inference with Jeffreys and matching priors as well as modified profile likelihoods for the cases where X and Y are normal or exponential random variables.

It is important to mention that the parameter θ may not be available in a closed form in many cases (see Azzalini and Chiogna, 2004 and Gupta and Brown, 2001 for an example of this). This makes difficult (if at all feasible) to find a reparameterisation involving θ, which complicates the use of the classical approach. In particular, the use of the profile likelihood might be difficult if this reparameterisation is not available (Díaz–Francés and Montoya, 2012). Alternative inferential approaches that overcome this difficulty are Bayesian inference, nonparametric estimation, and bootstrap; given that using these approaches it is possible to obtain bootstrap confidence intervals and credible intervals from the corresponding samples of $\hat{\theta}$ and θ, respectively (Baklizi and Eidous, 2006; Zhou, 2008; Rubio and Steel, 2012).

New interest has been focused on the estimation of θ in the case where X and Y are dependent random variables. For example Barbiero (2011) assumes that (X, Y) are jointly normally distributed; Rubio and Steel (2012) suppose that X and Y are marginally distributed as skewed scale mixture of normals and construct the corresponding joint distribution using a Gaussian Copula; Domma and Giordano (2012a) construct the joint distribution of (X, Y) using a Farlie-Gumbel-Morgenstern copula with marginal distributions belonging to the Burr system; Domma and Giordano (2012b) consider Dagum distributed marginals and construct their joint distribution using a Frank copula; among others (Nadarajah, 2005; Gupta et al., 2012). In these papers, the importance of taking the assumption of dependence between X and Y into account is illustrated using simulated and real data sets.

We propose a class of nonparametric estimators of θ for the case where (X, Y) are paired, possibly dependent, continuous random variables. This scenario is of interest given that paired observations are produced in many experimental designs (see e.g. Sprott, 2000 and Cox and Reid, 2000 for examples of this). The estimators proposed here are based on nonparametric estimators of the survival function and density function of $Y - X$. This approach avoids making distributional assumptions over (X, Y) and allows interval estimation of θ via nonparametric bootstrap. In addition, this method can be easily implemented in R using already existing packages. In Section 2 we introduce these estimators and discuss some of their properties. In Section 3 we present two examples, using simulated and real data, which illustrate the importance of accounting for pairing and dependence of the observations when conducting inference about θ.

2 Nonparametric estimators of θ

Let (X, Y) be a pair of continuous random variables. Let $(\mathbf{x}, \mathbf{y})$ be a sample from (X, Y) of size n and suppose that these observations are collected in couples (x_i, y_i), $i = 1, \dots, n$. Define the variable $Z = Y - X$ and the vector of differences $\mathbf{z} = \mathbf{y} - \mathbf{x}$. By definition, we have that

$$\theta = \mathbb{P}(Z > 0) = 1 - F_Z(0) = S_Z(0),$$

where F_Z and S_Z are the cumulative distribution function and the survival function of Z, respectively. If F_Z or S_Z are replaced by a nonparametric estimator, then we find an immediate connection between the nonparametric estimation of the cumulative distribution function (or the survival function) of Z and the nonparametric estimation of θ. Based on this, we propose the following algorithm for estimating θ.

Algorithm 1

```
1: Calculate the differences $\mathbf{z} = \mathbf{y} - \mathbf{x}$.
2: Using the sample $\mathbf{z}$ construct a nonparametric estimator $\hat{F}_Z$ of the distribution function of
   $Z$ and define the estimator $\hat{\theta} = 1 - \hat{F}_Z(0)$.
```

It is possible to define an alternative estimator of θ in Step 2 of Algorithm 1 by constructing a nonparametric estimator $\hat{f}_Z$ of the density of Z, based on the sample $\mathbf{z}$, and defining the estimator $\hat{\theta} = \int_0^\infty \hat{f}_Z(z)dz$. Several nonparametric estimators $\hat{F}_Z$ and $\hat{f}_Z$ can be considered for this purpose. For instance, kernel density estimators (Parzen, 1962), the empirical distribution function, shape-restricted density estimators (Cule et al., 2010) and recently proposed smoothed versions of these (Dümbeng and Rufibach, 2011; Rufibach, 2012). Note that the asymptotic properties of the estimator $\hat{\theta}$ are inherited from those of the estimator $\hat{F}_Z$ evaluated at 0. For example, if we use the empirical distribution function for estimating $\hat{F}_Z(0)$, then we have that $\hat{\theta} \overset{a.s.}{\to} \theta$ as $n \to \infty$. The use of nonparametric bootstrap on the sample $\mathbf{z}$ together with Algorithm 1 allows us to obtain a variety of bootstrap confidence intervals for θ (DiCiccio and Efron, 1996).

Note that this class of estimators avoids making assumptions on the distribution of (X, Y) and the sort of dependence between the variables X and Y. The relationship between these variables, which can be either dependent or independent, is implicitly included by modelling the differences between the observations which only requires a pairing of the observations.

3 Examples

In this section, we illustrate the implementation of the estimators proposed in Section 2. In the first example we use a sample simulated from a bivariate sinh-arcsinh distribution (Jones and Pewsey, 2009). As detailed in Jones and Pewsey (2009), this distribution contains parameters that control skewness, kurtosis and correlation of the marginals. This example illustrates the influence of the assumptions of pairing and dependence on the bootstrap distributions of $\hat{\theta}$ in terms of their location and spread. In the second example we use a real data set and show that not including the assumptions of pairing and dependence may lead to opposite conclusions about the relationship between X and Y.

In both examples, we consider the following 6 types of estimators of θ. Estimators based on Algorithm 1 with $\hat{\theta} = 1 - \hat{F}_Z(0)$: (1) The estimator "Kernel", based on a Gaussian kernel estimator of $\hat{F}_Z$; and (2) The estimator "ECDF", based on the empirical distribution function for estimating $\hat{F}_Z$. Estimators based on Algorithm 1 with $\hat{\theta} = \int_0^\infty \hat{f}_Z(z)dz$: (3) The estimator "MLE", where $\hat{f}_Z$ is the shape-restricted density estimator described in Cule et al. (2010); and (4) The estimator "SMLE", where $\hat{f}_Z$ is the smooth-shape-restricted density estimator proposed in Dümbeng and Rufibach (2011). For comparison purposes, we also consider two estimators based on the assumption of independence of X and Y: (5) The estimator "Independent" proposed in Baklizi and Eidous (2006), based on a Gaussian kernel estimator of the marginal densities of X and Y; and (6) The estimator "Paired", based on a Gaussian kernel estimator of the marginal densities of X and Y (Baklizi and Eidous, 2006) but taking the pairing of the observations into account in the bootstrap sampling.

Nonparametric density estimation is conducted using the R packages 'LogConcDEAD' (Cule et al., 2009) and 'logcondens' (Dümbeng and Rufibach, 2011). Bootstrap samples and bootstrap confidence intervals (Normal, Basic, Percentile and BCa) were obtained using the R packages 'boot' (Canty and Ripley, 2012) and 'simpleboot' (Peng, 2008). R source code for these examples is available upon request.

3.1 Simulated data

In this example we use a simulated sample of size $n = 100$ from a bivariate sinh-arcsinh distribution (Jones and Pewsey, 2009) with parameters $(\sigma_1, \sigma_2, \rho, \epsilon_1, \epsilon_2, \delta_1, \delta_2) = (1, 1, 0.75, 0, 1, 1, 2)$. Figure 1a shows a contour plot of the corresponding density. This is a complex scenario where the entries present departure from normality and correlation. The population correlation coefficient of this sample is 0.737 and the theoretical correlation is 0.743. The parameter θ in this family of distributions is not generally tractable. The theoretical value of θ, obtained by numerical integration, is 0.78. Figure 1b shows the bootstrap distribution of $\hat{\theta}$ using several nonparametric estimators. We can observe a considerable influence of the assumptions of pairing and dependence in the location and spread of the bootstrap distributions of $\hat{\theta}$. We can also notice the influence of these assumptions in the point estimators and bootstrap confidence intervals shown in Table 1. In this case, not including these assumptions leads to underestimating θ.

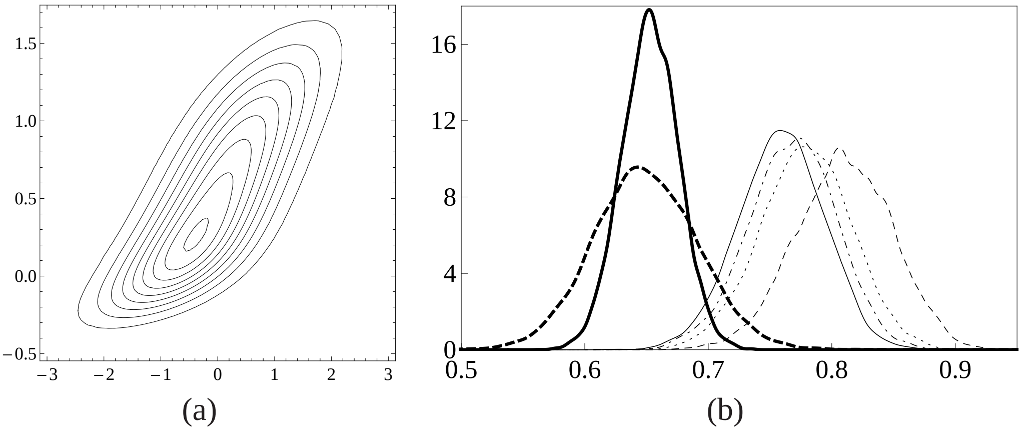

Figure 1: (a) Contour plot: sinh-arcsinh distribution; (b) Simulated data: bootstrap distributions of $\hat{\theta}$ using different estimators; “Independent” (bold-dashed line), “Paired” (bold line), “Kernel” (solid line), “ECDF” (dashed line), “MLE” (dotted line), “SMLE” (dotted–dashed line).

Estimator	$\hat{\theta}$	Normal	Basic	Percentile	BCa
Independent	0.65	$(0.560, 0.724)$	$(0.559, 0.723)$	$(0.568, 0.732)$	$(0.562, 0.727)$
Paired	0.65	$(0.606, 0.695)$	$(0.606, 0.696)$	$(0.607, 0.697)$	$(0.604, 0.694)$
Kernel	0.76	$(0.690, 0.825)$	$(0.692, 0.827)$	$(0.690, 0.824)$	$(0.684, 0.819)$
ECDF	0.81	$(0.734, 0.886)$	$(0.740, 0.890)$	$(0.730, 0.880)$	$(0.720, 0.870)$
MLE	0.78	$(0.707, 0.853)$	$(0.709, 0.854)$	$(0.705, 0.850)$	$(0.701, 0.847)$
SMLE	0.77	$(0.704, 0.844)$	$(0.707, 0.847)$	$(0.694, 0.835)$	$(0.694, 0.835)$

Table 1: Simulated data: Estimators and 95% bootstrap confidence intervals.

3.2 Real data

In this section we study the data set presented in Venkatraman and Begg (1996), which contains 72 lesion scores obtained using both a clinical scheme without a dermoscope (X Test), and a dermoscopic scoring scheme (Y Test). Their main interest is to assess the information provided by the use of the dermoscope. Here, we analyse the subset of 51 non-diseased patients (diagnosed using a biopsy) and compare the nonparametric inferences for θ obtained under three assumptions: independence, pairing and independence, and dependence of the tests using the estimators described in the introduction of this section. It is important to note that the population correlation coefficient of this sample is 0.794, which suggests that the entries are correlated.

Table 2 shows point estimators and four types of bootstrap confidence intervals of θ. Figure 2 shows the bootstrap distributions of $\hat{\theta}$ corresponding to the models described in Table 2. We can note a discrepancy of the point estimators under the assumptions of dependence and independence of the tests. Interval inference is also different; in the cases where pairing and dependence are not considered we can note that the value $\theta = 0.5$ is included in some of the bootstrap confidence intervals, leading to different conclusions about the relationship of the tests. This is in line with the conclusions in Rubio and Steel (2012) and emphasises the importance of the dependence and pairing assumptions.

Estimator	$\hat{\theta}$	Normal	Basic	Percentile	BCa
Independent	0.55	$(0.469, 0.678)$	$(0.467, 0.672)$	$(0.450, 0.656)$	$(0.474, 0.691)$
Paired	0.55	$(0.498, 0.597)$	$(0.497, 0.596)$	$(0.501, 0.601)$	$(0.499, 0.598)$
Kernel	0.63	$(0.5245, 0.737)$	$(0.525, 0.738)$	$(0.528, 0.741)$	$(0.519, 0.732)$
ECDF	0.69	$(0.559, 0.813)$	$(0.569, 0.823)$	$(0.549, 0.804)$	$(0.529, 0.784)$
MLE	0.65	$(0.543, 0.776)$	$(0.544, 0.776)$	$(0.532, 0.765)$	$(0.537, 0.768)$
SMLE	0.64	$(0.538, 0.756)$	$(0.539, 0.757)$	$(0.527, 0.744)$	$(0.533, 0.749)$

Table 2: Melanoma data: Estimators and 95% bootstrap confidence intervals.

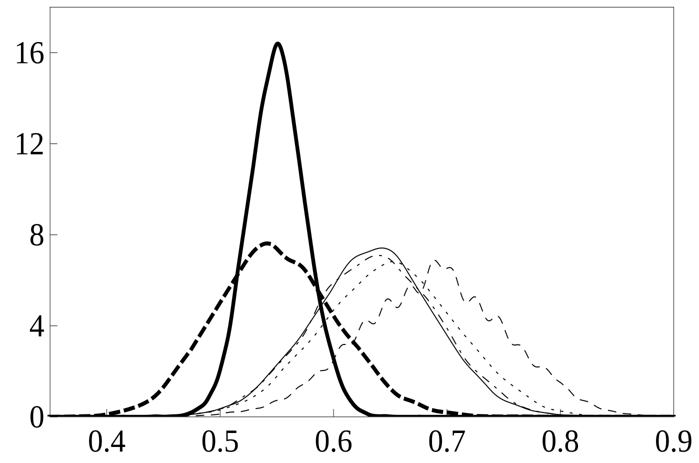

Figure 2: Melanoma data: bootstrap distributions of $\hat{\theta}$ using different estimators; "Independent" (bold-dashed line), "Paired" (bold line), "Kernel" (solid line), "ECDF" (dashed line), "MLE" (dotted line), "SMLE" (dotted–dashed line).

4 Discussion

We presented a class of nonparametric estimators for $\theta = P(X < Y)$ for the case of paired, possibly dependent, observations. This class of estimators avoids making assumptions on the distribution and the dependence structure of (X, Y), which are implicitly included in the estimation by modelling the differences of the observations. Confidence intervals for θ, based on these estimators, can be obtained using bootstrap methods which are easy to implement in R. It was illustrated, using a real data set, that not accounting for these assumptions might lead to opposite conclusions about $\theta = 0.5$, and consequently about the relationship between the variables X and Y.

A possible extension of this work consists of estimating θ in the context of censored and missing observations. The ideas presented here can be extended to these scenarios by using that

$$\theta = \int_{\mathbb{R}} \int_{-\infty}^{y} f_{X,Y}(x, y) dx dy,$$

and replacing the joint density $f_{X,Y}$ with a nonparametric density estimator. The use of kernel density estimators in these contexts has been studied, for example, in Titterington and Mill (1983) and Wells and Yeo (1996).

References

Azzalini, A. and Chiogna, M. (2004). Some results on the stress-strength model for skew-normal variates. *Metron* LXII: 315-326.

Baklizi, A. and Eidous, O. (2006). Nonparametric estimatrion of $P(X < Y)$ using kernel methods. *Metron* LXIV: 47–60.

Barbiero, A. (2011). Interval estimators for reliability: the bivariate normal case. *Journal of Applied Statistics*, doi: 10.1080/02664763.2011.602055.

Birnbaum, Z. M. (1956). On a use of the Mann–Whitney statistic, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Contributions to the Theory of Statistics, 13–17, University of California Press, Berkeley.

Canty, A. and Ripley, B. D. (2012). *boot: Bootstrap R (S–Plus) Functions.* R package version 1.3–5

Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments.* Chapman & Hall/CRC, Boca Raton, FL.

Cule, M., Gramacy, Robert and Samworth, R. (2009). LogConcDEAD: an R package for maximum likelihood estimation of a multivariate log–concave density . *Journal of Statistical Software* 29(2). URL http://www.jstatsoft.org/v29/i02/

Cule, M. L., Samworth, R. J. and Stewart, M. I. (2010), Maximum likelihood estimation of a multi–dimensional log–concave density. *Journal Royal Statistical Society* B 72: 545–600.

Díaz–Francés, E. and Montoya J. A. (2012). The simplicity of likelihood based inferences for $P(X < Y)$ and for the ratio of means in the exponential model. *Statistical Papers*, forthcoming.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science* 3: 189–228.

Domma, F. and Giordano, S. (2012a). A copula–based approach to account for dependence in stress–strength models. *Statistical Papers*, forthcoming.

Domma, F. and Giordano, S. (2012b). A stress-strength model with dependent variables to measure household financial fragility. *Statistical Methods and Applications*, forthcoming.

Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations Related to Univariate Log–Concave Density Estimation. *Journal of Statistical Software* 39: 1–28. URL http://www.jstatsoft.org/v39/i06/

Gupta, R. C. and Brown, N. (2001). Reliability studies of the skew–normal distribution and its application to a strength–stress model. *Communications in Statistics: Theory and Methods* 30: 2427-2445.

Gupta R. C., Ghitany, M. E. and Al–Mutairi, D. K. (2012). Estimation of reliability from a bivariate log–normal data. *Journal of Statistical Computation and Simulation*, forthcoming.

Jing, B. Y., Yuan, J. and Zhou, W. (2009). Jackknife empirical likelihood. *Journal of the American Statistical Association* 104, 1224–1232.

Jones, M. C. and Pewsey A. (2009). Sinh–arcsinh distributions. *Biometrika* 96: 761–780.

Kotz, S., Lumelskii, S. and Pensky, M. (2003). *The Stress–Strength Model and its Generalizations.* Theory and Applications. Singapore: World Scientific.

Montoya, J. A. (2008). *La verosimilitud perfil en la Inferencia Estadística.* PhD Thesis, Centro de Investigación en Matemáticas A. C., México.

Nadarajah, S. (2005). Reliability for some bivariate beta distributions. *Mathematical Problems in Engineering* 2: 101-111.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* 33: 1065–1076.

Peng, R. D. *simpleboot: Simple Bootstrap Routines.* R package version 1.1–3

Rubio, F. J. and Steel, M. F. J. (2011). Bayesian inference for $P(X < Y)$ using asymmetric dependent distributions. *Bayesian Analysis*, forthcoming.

Rufibach, K. (2012). A smooth ROC curve estimator based on log–concave density estimates. *International Journal of Biostatistics*, forthcoming.

Sprott, D. A. (2000). *Statistical Inference in Science.* Springer, New York.

Sun, D., Ghosh, M. and Basu, A. P. (1998). Bayesian analysis for a stress–strength system under noninformative priors. *The Canadian Journal of Statistics* 26: 323–332.

Titterington, D. M. and Mill, G. M. (1983). Kernel–based density estimates from incomplete data. *Journal of the Royal Statistical Society* B 45: 258–266.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution–free procedure for comparing operating characteristic curves from a paired experiment. *Biometrika* 83: 835–848.

Ventura, L. and Racugno, W. (2011). Recent advances on Bayesian inference for $P(X < Y)$. *Bayesian Analysis* 6: 1-18.

Wells, M. T. and Yeo, K. P. (1996). Density estimation with bivariate censored data. *Journal of the American Statistical Association* 436: 1566–1574.

Zhou, W. (2008). Statistical inference for $P(X < Y)$. *Statistics in Medicine* 27: 257–279.