A Unified Approach for Nonparametric Evaluation of Agreement in Method Comparison Studies

Pankaj K. Choudhary*

*University of Texas at Dallas, pankaj@utdallas.edu

A Unified Approach for Nonparametric Evaluation of Agreement in Method Comparison Studies*

Pankaj K. Choudhary

Abstract

We present a nonparametric methodology for evaluation of agreement between multiple methods of measurement of a continuous variable. Our approach is unified in that it can deal with any scalar measure of agreement currently available in the literature, and can incorporate repeated and unreplicated measurements, and balanced as well as unbalanced designs. Our key idea is to treat an agreement measure as a functional of the joint cumulative distribution function of the measurements from multiple methods. This measure is estimated nonparametrically by plugging-in a weighted empirical counterpart of the joint distribution function. The resulting estimator is shown to be asymptotically normal under some specified mild assumptions. A closed-form expression is provided for the asymptotic standard error of the estimator. This asymptotic normality is used to derive a large-sample distribution-free methodology for simultaneously comparing the multiple measurement methods. The small-sample performance of this methodology is investigated via simulation. The asymptotic efficiency of the proposed nonparametric estimator relative to the normality-based maximum likelihood estimator is also examined. The methodology is illustrated by applying it to a blood pressure data set involving repeated measurements from three measurement methods.

KEYWORDS: concordance correlation, multiple comparisons, repeated measurements, statistical functional, total deviation index, weighted empirical distribution function

---

1 Introduction

We consider the problem of agreement evaluation that arises in method comparison studies in health sciences research. These studies try to determine if $m\,(\geq 2)$ methods of measurement of a continuous clinical variable, such as blood pressure, cholesterol level, heart rate, etc., agree sufficiently well to be used interchangeably. A measurement method may be an instrument, a medical device, an assay, an observer or a measurement technique. The data in method comparison studies consist of one or more measurements by each measurement method on every experimental unit. Specifically, let X_{ijk}, $k = 1, \ldots, n_{ij}\,(\geq 1)$, $j = 1, \ldots, N$, $i = 1, \ldots, m$, denote the observed measurements, where X_{ijk} represents the kth replicate measurement on the jth experimental unit from the ith method. Here N is the number of units in the study.

Let θ be a measure of agreement between two measurement methods. This θ is a function of parameters of the bivariate distribution of measurements from the two methods and it quantifies the extent of agreement between the methods. The specific form of the function depends on the measure of agreement being used. We assume that θ is scalar and either a large or a small value for θ implies good agreement. In particular, let θ_{uv} be the value of θ that measures agreement between methods u and v, where (u, v) belongs to the index set $\mathcal{S}$ of (u, v) pairs, $u < v = 1, \ldots, m\,(\geq 2)$, which indicates the specific p pairs of measurement methods whose agreement evaluation is of interest. Thus, when all pairwise comparisons are of interest, $p = \binom{m}{2}$ and $\mathcal{S} = \{(u, v) : u < v = 1, \ldots, m\}$; and when comparisons with a reference method (say, method 1) are of interest, $p = m - 1$ and $\mathcal{S} = \{(1, v), v = 2, \ldots, m\}$. Moreover, when $m = 2$, we have $p = 1$ and $\mathcal{S} = \{(1, 2)\}$. Let $\boldsymbol{\theta}$ denote the p-vector with components θ_{uv}, $(u, v) \in \mathcal{S}$.

To discover the pairs of measurement methods that agree sufficiently well for interchangeable use, one performs multiple comparisons by computing simultaneous one-sided confidence intervals for the components of $\boldsymbol{\theta}$. Lower or upper confidence bounds are needed depending upon whether a large or a small value for θ implies good agreement. Frequently, however, two-sided intervals are also used in place of one-sided bounds. The goal of this article is to present a nonparametric methodology for computing these simultaneous intervals.

Several choices exist in the literature for a measure of agreement θ. They include limits of
agreement (Bland and Altman 1986), concordance correlation coefficient (CCC; Lin 1989), mean

squared deviation (MSD; Lin 2000), total deviation index (TDI; Lin 2000, and Choudhary and Na-
garaja 2007) and coverage probability (CP; Lin et al. 2002). A vast majority of the methodologies
currently available for performing inference on these agreement measures assume normality for the
measurements, see e.g., Bland and Altman (1986, 1999), Lin (1989), Lin (2000), Lin et al. (2002),
Carrasco and Jover (2003), Choudhary and Nagaraja (2007), Choudhary (2008), and Carstensen,
Simpson and Gurrin (2008). Sometimes nonparametric approaches (King and Chinchilli 2001a,
b, King, Chinchilli and Carrasco 2007, and Guo and Manatunga 2007) and approaches based
on generalized estimating equations (GEE; Barnhart and Williamson 2001, Barnhart, Song and
Haber 2005, and Lin, Hedayat and Wu 2007) are used as well. See Barnhart, Haber and Lin (2007)
for a review of the literature on the topic of agreement evaluation.

The authors such as Bland and Altman (1999) and Hawkins (2002) advocate the use of repli-
cated (or repeated) measurements in method comparison studies. When the measurements are
replicated, it is often the case that the resulting replicate measurements are unpaired (Chinchilli
et al. 1996) or that the design is unbalanced, i.e., not all n_{ij} are equal. But we are not aware
of any nonparametric methodology for agreement evaluation that can deal with such data; and
this is the primary motivation behind our work. The repeated measurements are said to be un-
paired when the measurements are replicated without any regard for timing of the measurements.
This scenario may occur, e.g., when the specimen of a subject is subsampled to yield multiple
measurements or when the repeated measurements are taken in a quick succession. Examples of
unpaired repeated measurements include the serum cholesterol data and the dietary intake data
of Chinchilli et al. (1996), and the blood pressure data of Bland and Altman (1999). These blood
pressure data are used later in this article for illustration. Further, an example of unbalanced
design is the cardiac output data of Bland and Altman (1999).

We focus on a nonparametric distribution-free paradigm in this article to avoid making as-
sumptions about the shape of the distribution of the measurements. Note that there are two

distinct types of dependencies in the measurements on an experimental unit — one is the depen-

dence among the repeated measurements from the same measurement method owing to a common

measurement method and a common experimental unit; and the other is the dependence among the measurements from different measurement methods owing to a common experimental unit. If the measurements are unreplicated, the first type of dependence does not exist and the standard nonparametric techniques designed for independently and identically distributed multivariate data (Lehmann 1998, Chapter 6) can be used for inference on an agreement measure. But these techniques need to be extended to deal with the repeated measurements data. So the novel contribution of this article is the development of a nonparametric methodology that makes use of the special dependence structure in the repeated measurements method comparison data for performing inference on an agreement measure. This methodology may be seen as an alternative to the usual parametric model-based approach that makes assumptions regarding the shape of the distribution of the measurements, which may unduly influence the resulting inference.

We treat an agreement measure as a *statistical functional* (Lehmann 1998, Chapter 6), i.e., a functional of the joint cumulative distribution function (cdf) of measurements from multiple methods and estimate the measure nonparametrically by plugging-in a weighted empirical counterpart of the cdf. The weights are used to take into account of the dependence. The resulting estimator is shown to be asymptotically normal under some specified mild assumptions. This result is used to derive an asymptotically distribution-free methodology for computing the desired simultaneous confidence intervals. The advantage of the statistical functional approach is that it enables us to present a unified methodology that can accommodate repeated and unreplicated measurements, balanced as well as unbalanced designs, multiple methods, and any scalar measure of agreement; including all the aforementioned measures except the limits of agreement, which uses two limits for measuring agreement.

Our nonparametric methodology is complementary to the existing methodologies for agreement evaluation with repeated measurements. Although the methodology of King et al. (2007a, b) is also nonparametric, but it is designed exclusively for CCC and it assumes that the repeated measurements are longitudinal and paired. The GEE-based methodology of Barnhart, Song and Haber (2005) is also designed only for CCC. The GEE-based methodology of Lin et al. (2007)

can accommodate many of the agreement measures listed above, but it assumes a balanced design and a two-way mixed-effects model for the data. Further, the methodology of Choudhary and Yin (2010) can be employed for inference on any scalar measure of agreement, but it assumes normality for measurements.

The rest of this article is organized as follows. Section 2 explains the proposed nonparametric methodology for computing simultaneous confidence intervals for the components of $\boldsymbol{\theta}$. The methodology is illustrated in Section 3 by applying it to a blood pressure data set. Results of a Monte Carlo simulation study are summarized in Section 4. Section 5 describes the theoretical underpinnings of the proposed methodology. Section 6 concludes with a discussion.

2   Methodology for nonparametric confidence intervals

Let the m-vector $\mathbf{X} = (X_1, \ldots, X_m)$ denote the measurements from m methods on a randomly selected experimental unit from the population. Also, let X_i' be a replicate of X_i, $i = 1, \ldots, m$, from the same unit. Next, let F be the joint cdf of $\mathbf{X}$ and the compact set $\mathcal{X} \subseteq \overline{\mathbb{R}}^m$ be the support of $\mathbf{X}$, where $\overline{\mathbb{R}}$ is the extended real line $[-\infty, \infty]$. Recall that X_{ijk}, $k = 1, \ldots, n_{ij}$, $j = 1, \ldots, N$, $i = 1, \ldots, m$, denote the observed measurements. We now make three basic assumptions.

A1. The cdf $F(\mathbf{x})$ is continuous in $\mathbf{x} = (x_1, \ldots, x_m) \in \mathcal{X}$.

A2. The measurements on different experimental units are independent.

A3. For each $j = 1, \ldots, N$, the $\prod_{i=1}^m n_{ij}$ possible m-tuples formed by the measurements on the jth experimental unit, i.e., $\{(X_{1jk_1}, X_{2jk_2}, \ldots, X_{mjk_m}),\ k_i = 1, \ldots, n_{ij},\ i = 1, \ldots, m\}$, are identically distributed as $\mathbf{X}$.

The assumption A3 implies that each X_{ijk} is identically distributed as X_i.

2.1   The agreement measure as a statistical functional

Consider θ_{uv}, the value of $\boldsymbol{\theta}$ that measures agreement between methods u and v, $(u, v) \in \mathcal{S}$. By definition, θ_{uv} is a characteristic of the joint cdf F_{uv} of (X_u, X_v), or more generally, a characteristic

of the joint cdf F of $\mathbf{X}$. Hence θ_{uv} is a statistical functional, say $\theta_{uv} = h_{uv}(F)$, where h_{uv} is a known

real-valued function defined over a class $\mathcal{F}$ of multivariate cdfs on $\mathcal{X}$ for which θ_{uv} is well-defined.

Take, for example, four measures of agreement mentioned in Section 1, namely, MSD (Lin 2000),

CP (Lin et al. 2002), TDI (Lin 2000) and CCC (Lin 1989). They are defined as follows:

$$MSD_{uv} = E(X_u - X_v)^2,$$

$$CCC_{uv} = \frac{2cov(X_u, X_v)}{\{E(X_u) - E(X_v)\}^2 + var(X_u) + var(X_v)} = \frac{2\{E(X_u X_v) - E(X_u)E(X_v)\}}{E(X_u^2) + E(X_v^2) - 2E(X_u)E(X_v)},$$

$$CP_{uv} = G_{uv}(\delta), \text{ for a given small } \delta > 0,$$

$$TDI_{uv} = G_{uv}^{-1}(\nu), \text{ for a given large probability } \nu \in (0, 1), \tag{1}$$

where $G_{uv} = \text{cdf of } |X_u - X_v|$ and $G_{uv}^{-1}(\nu) = \inf\{t : G_{uv}(t) \geq \nu\}$ is the νth quantile of $|X_u - X_v|$.

The measures MSD, CP and TDI are positive, and CCC lies in $(-1, 1)$. In case of MSD and TDI,

small values for the measures imply good agreement, whereas in case of CP and TDI, large values

for the measures imply good agreement. These measures can be written as statistical functionals

in the following manner:

$$MSD_{uv}(F) = \int_{\mathcal{X}} (x_u - x_v)^2 \, dF(\mathbf{x}),$$

$$CCC_{uv}(F) = \frac{2\left\{\int_{\mathcal{X}} x_u x_v \, dF(\mathbf{x}) - \int_{\mathcal{X}} x_u \, dF(\mathbf{x}) \cdot \int_{\mathcal{X}} x_v \, dF(\mathbf{x})\right\}}{\int_{\mathcal{X}} x_u^2 \, dF(\mathbf{x}) + \int_{\mathcal{X}} x_v^2 \, dF(\mathbf{x}) - 2\int_{\mathcal{X}} x_u \, dF(\mathbf{x}) \cdot \int_{\mathcal{X}} x_v \, dF(\mathbf{x})},$$

$$CP_{uv}(F) = \int_{\mathcal{X}} I\big(|x_u - x_v| \leq \delta\big) \, dF(\mathbf{x}), \text{ for a given } \delta > 0,$$

$$TDI_{uv}(F) = \inf\left\{t : \int_{\mathcal{X}} I\big(|x_u - x_v| \leq t\big) \, dF(\mathbf{x}) \geq \nu\right\}, \text{ for a given } \nu \in (0, 1), \tag{2}$$

where $I(A)$ is the indicator of the event A.

Since $\theta_{uv} = h_{uv}(F)$ is a statistical functional, the p-vector $\boldsymbol{\theta}$ with components θ_{uv}, $(u, v) \in \mathcal{S}$,

is also a statistical functional $\boldsymbol{\theta} = \mathbf{h}(F)$, where $\mathbf{h} : \mathcal{F} \mapsto \mathbb{R}^p$ is a known function, which essentially

stacks h_{uv}, $(u, v) \in \mathcal{S}$, into a p-vector. Let $\hat{F}$ be the empirical counterpart of F. Then, the plug-

in estimator $\hat{\boldsymbol{\theta}} = \mathbf{h}(\hat{F})$, the p-vector with components $\hat{\theta}_{uv} = h_{uv}(\hat{F})$, $(u, v) \in \mathcal{S}$, is the natural

nonparametric estimator of $\boldsymbol{\theta}$. The next section describes $\hat{F}$.

2.2 Empirical cdf $\hat{F}$

When the measurements are repeated, unlike the case of unreplicated measurements, the estimator of F is not uniquely defined. We focus on a weighted empirical cdf of the form:

$$\hat{F}(\mathbf{x}) = \sum_{j=1}^{N} w(N, \mathbf{n}_j) \left\{ \sum_{k_1=1}^{n_{1j}} \cdots \sum_{k_m=1}^{n_{mj}} I(X_{1jk_1} \leq x_1, \ldots, X_{mjk_m} \leq x_m) \right\}, \tag{3}$$

where w is a weight function; $\mathbf{n}_j$ is the m-vector $(n_{1j}, \ldots, n_{mj})$; and $\{(X_{1jk_1}, \ldots, X_{mjk_m}), k_i = 1, \ldots, n_{ij}, i = 1, \ldots, m\}$ are the $\prod_{i=1}^{m} n_{ij}$ possible m-tuples formed by the measurements on the jth unit. The function w is assumed to satisfy $\sum_{j=1}^{N} \left\{ \prod_{i=1}^{m} n_{ij} \right\} w(N, \mathbf{n}_j) = 1$, ensuring unbiasedness of $\hat{F}$. The weights in (3) are free of $\mathbf{x}$ and depend on unit j only through the number of replications on the unit. When the design this balanced, i.e., all $n_{ij} = n$, this unbiasedness condition implies that $w(N, \mathbf{n}_j) = 1/(n^m N)$. Thus, the weights are unique in this case. Further, when the measurements are unreplicated, i.e., $n = 1$; $w(N, \mathbf{n}_j) = 1/N$ and $\hat{F}$ reduces to the usual empirical cdf. Under the empirical distribution (3), the joint and marginal probability mass functions of X_u and X_v, $(u, v) \in \mathcal{S}$, are:

$$P_{\hat{F}}(X_u = x_{ujk_u}, X_v = x_{vjk_v}) = \{w(N, \mathbf{n}_j)/(n_{uj} n_{vj})\} \prod_{i=1}^{m} n_{ij},$$

$$P_{\hat{F}}(X_u = x_{ujk_u}) = \{w(N, \mathbf{n}_j)/n_{uj}\} \prod_{i=1}^{m} n_{ij}, \quad P_{\hat{F}}(X_v = x_{vjk_v}) = \{w(N, \mathbf{n}_j)/n_{vj}\} \prod_{i=1}^{m} n_{ij}, \tag{4}$$

where $k_u = 1, \ldots, n_{uj}$, $k_v = 1, \ldots, n_{vj}$, $j = 1, \ldots, N$.

Following Olsson and Rootzén (1996), who consider the estimation of a univariate cdf with repeated measurements data, it is possible to find the optimal weight function that makes $\hat{F}$ the minimum variance unbiased estimator of F. This optimal function involves $\mathbf{x}$ and unknown covariances of the indicators in (3). But unfortunately the resulting $\hat{F}$ may not be a valid cdf since it may not be non-decreasing in $\mathbf{x}$. As a consequence, some of the observed measurements may have negative "probabilities" under the empirical distribution. Since this would create difficulty in deriving estimators of agreement measures, we do not consider the optimal weight function. Nevertheless, it can be shown that this optimal weight function reduces to

$$w_1(N, \mathbf{n}_j) = \frac{1}{N \prod_{i=1}^{m} n_{ij}} \quad \text{and} \quad w_2(N, \mathbf{n}_j) = \frac{1}{\sum_{j=1}^{N} \prod_{i=1}^{m} n_{ij}}, \tag{5}$$

respectively when the indicators in (3) have correlation one and zero. Both w_1 and w_2 are free of

$\mathbf{x}$ and can be considered as extreme special cases of the weight function w in (3). The function w_1 assigns $1/N$ weight to each unit in the study and distributes it equally over all m-tuples from this unit, whereas the function w_2 assigns equal weight to every m-tuple in the data. All three functions, w_1, w_2 and the optimal weight function, are identical when the design is balanced. The simulation study in Section 4 provides some guidance on how to choose between w_1 and w_2 for unbalanced designs.

2.3   Simultaneous confidence intervals for $\boldsymbol{\theta} = \mathbf{h}(F)$

We now explain the proposed methodology for computing simultaneous confidence intervals for the components of $\boldsymbol{\theta}$. The technical details underlying this methodology are postponed to Section 5. When N is large, under certain assumptions, the plug-in estimator $\hat{\boldsymbol{\theta}} = \mathbf{h}(\hat{F})$ approximately follows a $\mathcal{N}_p(\boldsymbol{\theta}, \Sigma/N)$ distribution, where Σ is given by (14). This covariance matrix is defined in terms of moments of the *influence function* of θ_{uv}, say, $L_{uv}(x_u, x_v) \equiv L_{uv}(\mathbf{x}, F)$, $(u, v) \in \mathcal{S}$. The influence function of a statistical functional measures the rate at which the functional changes when F is contaminated by a small probability of the contamination $\mathbf{x}$. It plays an important role in the asymptotic theory of nonparametric and robust estimators (Lehmann 1998, Chapter 6). To define the influence function, let $\delta_{\mathbf{x}}$ be the cdf of an m-variate distribution that assigns probability one to the point $\mathbf{x} \in \mathbb{R}^m$. Then, $L_{uv}(x_u, x_v) = \frac{d}{d\epsilon} h_{uv} \{(1 - \epsilon)F + \epsilon\delta_{\mathbf{x}}\} |_{\epsilon=0}$ and it satisfies $\int_{\mathcal{X}} L_{uv}(x_u, x_v) dF(\mathbf{x}) = 0$. Let $\hat{L}_{uv}(x_u, x_v) \equiv L_{uv}(\mathbf{x}, \hat{F})$ be the empirical counterpart of the influence function. The sample moments of this empirical influence function are used to construct an estimator $\hat{\Sigma}$ of Σ in Section 5.2.

The asymptotic normality of $\hat{\boldsymbol{\theta}}$ suggests the following confidence intervals for $\boldsymbol{\theta}$:

$$\text{Upper bounds for } \theta_{uv}: \ \hat{\theta}_{uv} + \hat{c}_{1-\alpha,p}\, \hat{\sigma}_{uv}/N^{1/2},$$

$$\text{Lower bounds for } \theta_{uv}: \ \hat{\theta}_{uv} - \hat{c}_{1-\alpha,p}\, \hat{\sigma}_{uv}/N^{1/2},$$

$$\text{Two-sided intervals for } \theta_{uv}: \ \hat{\theta}_{uv} \pm \hat{d}_{1-\alpha,p}\, \hat{\sigma}_{uv}/N^{1/2}, \tag{6}$$

7

for $(u, v) \in \mathcal{S}$, where $\hat{\sigma}_{uv}^2$ is a diagonal element of $\hat{\Sigma}$, and $\hat{c}_{1-\alpha,p}$ and $\hat{d}_{1-\alpha,p}$ are critical points that

ensure approximately $(1 - \alpha)$ simultaneous coverage probability when N is large. To define the

critical points, consider a p-vector with elements Z_{uv}, $(u, v) \in \mathcal{S}$, whose joint distribution is normal with mean zero and covariance matrix equal to the correlation matrix corresponding to Σ. Let $c_{1-\alpha,p}$ and $d_{1-\alpha,p}$ be $(1-\alpha)$th percentiles of $\max_{(u,v)\in\mathcal{S}} Z_{uv}$ and $\max_{(u,v)\in\mathcal{S}} |Z_{uv}|$, respectively. Then, $\hat{c}_{1-\alpha,p}$ and $\hat{d}_{1-\alpha,p}$ are estimates of $c_{1-\alpha,p}$ and $d_{1-\alpha,p}$ obtained by replacing Σ in their definitions with $\hat{\Sigma}$. The validity of these intervals is established in Section 5.3.

Note that when $m = 2$; $p = 1$ and $c_{1-\alpha,1}$ and $d_{1-\alpha,1}$ are simply the $(1 - \alpha)$th and $(1 - \alpha/2)$th percentiles of a standard normal distribution. In practice, the critical points $\hat{c}_{1-\alpha,p}$ and $\hat{d}_{1-\alpha,p}$ can be computed using the method of Hothorn, Bretz and Westfall (2008), which is implemented in their `multcomp` package for the statistical software `R` (R Development Core Team 2009).

Often the small-sample properties of the intervals in (6) can be improved by first applying a transformation to the agreement measure, computing the intervals on the transformed scale, and then applying the inverse transformation to get the intervals on the original scale. In particular, the log transformation in case of MSD and TDI, the Fisher's z-transformation in case of CCC and the logit transformation in case of CP tend to work well (Lin et al. 2002).

To summarize, the proposed methodology for computing approximate $(1 - \alpha)$ simultaneous confidence intervals for $\theta_{uv}, (u, v) \in \mathcal{S}$ is as follows:

1. Verify that the assumption A8, given in Section 5.1, holds for θ_{uv}. This assumption is crucial for the asymptotic normality of the estimated agreement measures.

2. Estimate F using $\hat{F}$, given by (3).

3. Compute $\hat{\theta}_{uv} = h_{uv}(\hat{F})$ using the empirical distribution (4) of (X_u, X_v) under $\hat{F}$.

4. Find the influence function $L_{uv}(x_u, x_v)$ and compute its empirical counterpart $\hat{L}_{uv}(x_u, x_v)$. The influence function is needed to get $\hat{\Sigma}$.

5. Compute $\hat{\Sigma}$ as described in Section 5.2.

6. Compute the critical point $\hat{c}_{1-\alpha,p}$ or $\hat{d}_{1-\alpha,p}$ and use the intervals in (6).

2.4 The four agreement measures

We now return to the four specific measures introduced in (2), namely, MSD, CCC, CP and TDI.

Using the empirical joint distribution (4) of (X_u, X_v) under $\hat{F}$, their plug-in estimators are:

$$\widehat{MSD}_{uv} = \sum_{j=1}^{N} \{w(N, \mathbf{n}_j)/(n_{uj}n_{vj})\} \left\{\prod_{i=1}^{m} n_{ij}\right\} \sum_{k_u=1}^{n_{uj}} \sum_{k_v=1}^{n_{vj}} (X_{ujk_u} - X_{vjk_v})^2,$$

$$\widehat{CCC}_{uv} = \frac{2\{E_{\hat{F}}(X_u X_v) - E_{\hat{F}}(X_u)E_{\hat{F}}(X_v)\}}{E_{\hat{F}}(X_u^2) + E_{\hat{F}}(X_v^2) - 2E_{\hat{F}}(X_u)E_{\hat{F}}(X_v)},$$

$$\widehat{CP}_{uv} = \hat{G}_{uv}(\delta), \quad \widehat{TDI}_{uv} = \hat{G}_{uv}^{-1}(\nu) = \inf\{t : \hat{G}_{uv}(t) \geq \nu\}, \quad \text{where}$$

$$\hat{G}_{uv}(t) = \sum_{j=1}^{N} \{w(N, \mathbf{n}_j)/(n_{uj}n_{vj})\} \left\{\prod_{i=1}^{m} n_{ij}\right\} \sum_{k_u=1}^{n_{uj}} \sum_{k_v=1}^{n_{vj}} I(|X_{ujk_u} - X_{vjk_v}| \leq t), \quad t > 0, \quad (7)$$

and $\delta > 0$ and $\nu \in (0, 1)$ are specified. Moreover,

$$E_{\hat{F}}(X_u X_v) = \sum_{j=1}^{N} \{w(N, \mathbf{n}_j)/(n_{uj}n_{vj})\} \left\{\prod_{i=1}^{m} n_{ij}\right\} \sum_{k_u=1}^{n_{uj}} \sum_{k_v=1}^{n_{vj}} X_{ujk_u} X_{vjk_v},$$

$$E_{\hat{F}}(X_u^a) = \sum_{j=1}^{N} \{w(N, \mathbf{n}_j)/n_{uj}\} \left\{\prod_{i=1}^{m} n_{ij}\right\} \sum_{k_u=1}^{n_{uj}} X_{ujk_u}^a, \quad a = 1, 2,$$

$$E_{\hat{F}}(X_v^a) = \sum_{j=1}^{N} \{w(N, \mathbf{n}_j)/n_{vj}\} \left\{\prod_{i=1}^{m} n_{ij}\right\} \sum_{k_v=1}^{n_{vj}} X_{vjk_v}^a, \quad a = 1, 2.$$

To use the proposed confidence interval methodology for these measures, we need to verify the assumption A8 for them and find their influence functions. The assumption is verified in Section 5.4 and the influence functions $L_{uv}(x_u, x_v)$ are as follows.

$$MSD_{uv}(F): \ (x_u - x_v)^2 - MSD_{uv}(F),$$

$$CCC_{uv}(F): \ (A_{u,2} + A_{v,2} - 2A_{u,1}A_{v,1})^{-1} \big[2(CCC_{uv}(F) - 1)\{(x_u - A_{u,1})A_{v,1}$$

$$+(x_v - A_{v,1})A_{u,1}\} + 2(x_u x_v - A_{uv}) - CCC_{uv}(F)\{(x_u^2 - A_{u,2}) + (x_v^2 - A_{v,2})\}\big],$$

$$CP_{uv}(F): \ I(|x_u - x_v| \leq \delta) - CP_{uv}(F), \quad \text{for a given } \delta > 0,$$

$$TDI_{uv}(F): \ \{g_{uv}(TDI_{uv})\}^{-1} \{-I(|x_u - x_v| \leq TDI_{uv}) + G_{uv}(TDI_{uv})\}, \quad (8)$$

where $A_{u,a} = \int_{\mathcal{X}} x_u^a \, dF(\mathbf{x})$, $A_{v,a} = \int_{\mathcal{X}} x_v^a \, dF(\mathbf{x})$, $a = 1, 2$, and $A_{uv} = \int_{\mathcal{X}} x_u x_v dF(\mathbf{x})$; and $g_{uv}(TDI_{uv})$ is the derivative of the cdf G_{uv} at TDI_{uv}.

Note that in case of TDI, the computation of $\hat{\Sigma}$ using the influence function involves estimation of the densities $g_{uv}(TDI_{uv})$ in the tails. Unfortunately these density estimates are generally not stable unless N is quite large. So, a preferable alternative is to use the following simultaneous confidence intervals for TDI_{uv} that avoid the density estimation:

$$\text{Upper bounds:} \quad \hat{G}_{uv}^{-1}\left(\nu + \hat{c}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2}\right),$$

$$\text{Two-sided intervals:} \quad \hat{G}_{uv}^{-1}\left(\nu \pm \hat{d}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2}\right), \tag{9}$$

for $(u,v) \in \mathcal{S}$, where the critical points $\hat{c}_{1-\alpha,p}$ and $\hat{d}_{1-\alpha,p}$ are obtained as in (6) by taking $\theta_{uv} = G_{uv}(TDI_{uv})$, $\hat{\theta}_{uv} = \hat{G}_{uv}(\widehat{TDI}_{uv})$ and the influence function of θ_{uv} as $L_{uv}(x_u, x_v) = I(|x_u - x_v| \leq TDI_{uv}) - G_{uv}(TDI_{uv})$. Here $\hat{G}_{uv}(\widehat{TDI}_{uv})$ is obtained by simply substituting $\widehat{TDI}_{uv}$ for δ in the expression for $\widehat{CP}_{uv}$ given in (7).

3 Application

Consider the blood pressure data of Bland and Altman (1999). This data set has 85 subjects and on each subject three replicate measurements (in mmHg) of systolic blood pressure are made in quick succession by each of two experienced observers J and R (say, methods 1 and 2) using a sphygmomanometer and by a semi-automatic blood pressure monitor S (say, method 3). The interest is in simultaneously evaluating the extent of agreement between the three pairs of methods — (J, R), (J, S) and (R, S). Here we consider only two agreement measures — CCC and TDI with $\nu = 0.90$, and their one-sided bounds. The other measures can be handled in a similar manner.

The standard normality-based approach is to model these data as a mixed-effects model,

$$X_{ijk} = \mu_i + b_{ij} + \epsilon_{ijk}, \quad i = 1, 2, 3 \,(= m), \;\; j = 1, \ldots, 85 \,(= N), \;\; k = 1, 2, 3 \,(= n), \tag{10}$$

where X_{ijk} is the kth repeated measurement on jth subject from the ith measurement method; μ_i is the fixed effect of the ith measurement method; b_{ij} is the random effect of jth subject on ith measurement method; and ϵ_{ijk} is the error term. It is assumed that the interaction effects $(b_{1j}, b_{2j}, b_{3j}) \sim$ independent $\mathcal{N}_3(0, \Psi)$ distributions, with Ψ as an unstructured covariance matrix;

$\epsilon_{ijk} \sim$ independent $\mathcal{N}(0, \sigma_i^2)$ distributions; and the errors and the random effects are mutually independent.

Figure 1 presents the box plots of the resulting standardized residuals and standardized estimates of random effects when this model is fitted via maximum likelihood (ML) in `R` (R Development Core Team 2009) using the `nlme` package of Pinheiro et al. (2009). Since there is evidence of heavytailedness in the residuals and skewness in the random effects, the estimates and simultaneous bounds for agreement measures based on this model may not be accurate.

We now summarize the results of the proposed nonparametric analysis. The weights used in the empirical cdf $\hat{F}$ in (3) equal $w(N, \mathbf{n}_j) = 1/(n^3 N) = 1/(27 * 85)$ for all j, as the design is balanced. The estimated (mean, standard deviation) of measurements from methods J, R and S computed using the empirical distribution (4) are $(127.4, 31.0)$, $(127.3, 30.7)$ and $(143.0, 32.5)$, respectively. In addition, the estimated correlation between measurements from method pairs (J, R), (J, S) and (R, S) are 0.97, 0.79 and 0.79, respectively. Thus, the measurements from methods J and R have practically the same means and variances and their correlation is very high. On the other hand, the measurements from method S differ by those from methods J and R by about 16 mmHg on average. Moreover, the measurements from method S have somewhat higher variability than the other two methods and the correlation between them is relatively low.

Table 1 presents estimates, standard errors and 95% simultaneous lower bounds for CCC computed using (6). It also presents estimates and 95% simultaneous upper bounds for TDI using (9). The standard errors of TDI estimates are not presented as they are not needed for the bounds (9) that avoid density estimation. The CCC bounds are computed by first applying the Fisher's z-transformation to CCC. The critical point $\hat{c}_{0.95,3}$ equals -1.99 in case of CCC and 1.93 in case of TDI. Using the TDI bounds, we can conclude that 90% of the measurement differences between J and R, J and S, and R and S are estimated to lie between ±14, ±54 and ±53 mmHg, respectively. If one takes 15 mmHg as the margin of acceptable differences in blood pressure measurements, then the agreement between J and R is inferred to be acceptable, whereas the agreement between J and S, and R and S are inferred to be unacceptable. Moreover, J and S, and R and S appear to have comparable extent of agreement. The same conclusion is reached on the basis of CCC lower bounds. It is also evident that a substantial mean difference and a relatively low correlation is the

cause of unacceptable agreement between methods J and S, and R and S.

For comparison, we now report the results of the analysis assuming the mixed-effects model (10) for the data. Interestingly, the ML estimates of means, variances and correlations of measurements from the three methods are identical to the nonparametric estimates reported earlier. The simultaneous bounds for CCC and TDI, obtained using the methodology of Choudhary and Yin (2010) are also presented in Table 1. The nonparametric and the model-based estimates of CCC are identical. The two approaches also produce practically the same estimates and bounds for both CCC and TDI in case of (J, R) method pair, but the approaches do differ for method pairs (J, S) and (R, S). In particular, the parametric bounds overestimate the extent of agreement between these two method pairs.

4 Simulation study

4.1 Finite-sample coverage probability

To evaluate the finite-sample coverage probabilities of the proposed nonparametric TDI and CCC bounds, we simulate data from the model,

$$X_{ijk} = \mu_i + b_{ij} + \epsilon_{ijk}, \quad i = 1, 2, 3 \, (= m), \; j = 1, \ldots, N, \; k = 1, \ldots, n, \tag{11}$$

where μ_i, b_{ij} and ϵ_{ijk} are as defined in (10). This model is similar to the one used for the blood pressure data except that we assume a multivariate skew-t distribution with location vector $\mathbf{0}$ and scale matrix Ψ for (b_{1j}, b_{2j}, b_{3j}) and a univariate skew-t distribution with location zero and scale σ_i^2 for ϵ_{ijk}. Skew-t distributions (Azzalini and Capitanio 2003) are a generalization of the normal distribution that have a shape parameter to regulate skewness and a degrees-of-freedom parameter to control heavytailedness. Two combinations of these parameters are considered. This first has zero as the shape parameter and infinity as the degrees of freedom, leading to the usual normal-based mixed-effects model, and the second has 5 as both the shape parameter and the degrees of

freedom. The other model parameters are taken to be

$$\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 127 \\ 127 \\ 143 \end{bmatrix}, \quad \Psi = \begin{bmatrix} 900 & 891 & 772 \\ 891 & 900 & 772 \\ 772 & 772 & 961 \end{bmatrix}, \quad \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 6 \\ 9 \end{bmatrix}.$$

These values are essentially the rounded estimates from the blood pressure data. The computations are programmed in R (R Development Core Team 2009).

Table 2 summarizes the estimated simultaneous coverage probabilities of TDI and CCC bounds for $N = 30, 60, 100$; $n = 1, 2, 3, 4$; $1 - \alpha = 0.95$; and $\nu = 0.90$ for TDI. These summaries are based on 2500 samples from the model (11). When $n = 1$, b_{ij} in this model is replaced by b_{1j} to make the model identifiable. Further, the CCC lower bounds are computed using (6) by first applying the Fisher's z-transformation and the TDI lower bounds are computed using (9). In case of CCC, the estimated coverage probabilities tend to be close to 0.95 except when $\{N = 30, n = 1\}$ and the distribution is skew-t. In case of TDI, the coverage probabilities tend to be close to 0.95 when $\{N \geq 60, n \geq 2\}$, more than 0.95 when $\{N \geq 60, n = 1\}$, and less than 0.95 when $N = 30$.

Oftentimes in practice, the use of studentized bootstrap (Davison and Hinkley 1997) to compute critical points leads to more accurate confidence intervals than the standard normality-based critical points. However, in additional simulations (not presented here), we observed that the coverage probabilities of the bootstrap confidence bounds were closer to 0.95 than the bounds in (6) only in the case of CCC with $\{N = 30, n = 1\}$. In all other cases, the bootstrap bounds were quite conservative.

4.2 Asymptotic relative efficiency

We now study the asymptotic efficiency of the nonparametric estimators relative to the ML estimators. First, we consider balanced designs and focus on two special cases of model (11) with $m = 2$, namely, the normal model and the skew-t model with 5 as both the shape parameter and the degrees of freedom. The attention is restricted to two combinations of the remaining model parameters. One is the "high agreement" combination consisting of $\{(\mu_1, \mu_2) =$

13

$(127, 127), (\sigma_1, \sigma_2) = (6, 6), \Psi = \Psi_1\}$ and the other is the "low agreement" combination consisting

of $\{(\mu_1, \mu_2) = (127, 143), (\sigma_1, \sigma_2) = (6, 9), \Psi = \Psi_2\}$, with

$$\Psi_1 = \begin{bmatrix} 900 & 891 \\ 891 & 900 \end{bmatrix}, \ \Psi_2 = \begin{bmatrix} 900 & 772 \\ 772 & 961 \end{bmatrix}.$$

Table 3 reports the approximate asymptotic relative efficiency (ARE) of the nonparametric estimators of CCC and TDI with $\nu = 0.90$ with respect to their ML counterparts obtained by assuming the aforementioned normal model. This relative efficiency is the ratio of the mean squared errors of the nonparametric and the ML estimators with $N = 500$. They are based on 2500 Monte Carlo repetitions, and we take $n = 1, 2, 3, 4$ for this computation. When the true model is normal, the nonparametric estimators are not expected to be more efficient than the ML estimators as the latter are asymptotically efficient. Gain in efficiency, however, is expected when the true model is skew-t. In case of CCC, there is no practical difference between its nonparametric and ML estimators and hence they have the same efficiency for both models. On the other hand, the nonparametric estimator of TDI loses between 14-60% efficiency over the ML estimator when the true model is normal, whereas it gains between 56-74% efficiency when the true model is skew-t.

Next, we consider the case of unbalanced designs. The previous model is now modified so that 25% of experimental units have q replications from each measurement method, where $q = 1, 2, 3, 4$. The relative efficiencies now depend on which weight function w is used in the estimation. Table 3 presents these efficiencies for two weight functions w_1 and w_2, given by (5). The approximate w_2 vs. w_1 AREs obtained by taking the ratios of the corresponding nonparametric vs. ML AREs are also presented. In case of CCC, w_1 is always better than w_2. In case of TDI, however, w_1 is better than w_2 when the agreement is low, whereas the converse is true when the agreement is high. Further, there is a minor loss of efficiency of the nonparametric estimator of CCC with weight w_1 relative to the ML estimator. This indicates that, unlike the case of balanced designs, the two estimators of CCC may not be the same when the design is unbalanced. In case of TDI, the better nonparametric estimator loses about 30-40% efficiency over the ML estimator when the true model is normal, whereas it gains between 50-60% efficiency when the true model is skew-t. These results suggest that whether the weight function w_1 or w_2 will lead to a more efficient nonparametric

estimator will depend on the agreement measure of interest and the parameter values.

The w_2 vs. w_1 ARE can also be computed as the ratio of the asymptotic variances — (σ_{12}^2 with $w = w_2)/(\sigma_{12}^2$ with $w = w_1$). Under the assumed design, σ_{12}^2 in (14) can be simplified as

$$\sigma_{12}^2 = \frac{1}{4} \sum_{q=1}^{4} w^{*2}(q,q) \, q^2 \big[E\{L_{12}^2(X_1, X_2)\} + (q-1)E\{L_{12}(X_1, X_2)L_{12}(X_1', X_2)\}$$

$$+ (q-1)E\{L_{12}(X_1, X_2)L_{12}(X_1, X_2')\} + (q-1)^2 E\{L_{12}(X_1, X_2)L_{12}(X_1', X_2')\}\big].$$

Moreover, $w^*(q,q)$ equals $1/q^2$ if $w = w_1$ and $4/30$ if $w = w_2$. The moments in σ_{12}^2 can be approximated via Monte Carlo. The values of this ARE, also presented in Table 3, match closely with the above AREs computed as the ratios of mean squared errors with $N = 500$.

5 Technical details

In this section, we describe the theoretical justification for the confidence interval methodology proposed in Section 2. First, we make the following additional assumptions.

A4. For each $j = 1, \ldots, N$, the joint distribution of any two m-tuples $(X_{1jk_1}, \ldots, X_{mjk_m})$ and $(X_{1jl_1}, \ldots, X_{mjl_m})$ from the jth unit is the same as that of $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_m)$, where for $k_i, l_i = 1, \ldots, n_{ij}$, $i = 1, \ldots, m$,

$$Y_i = \begin{cases} X_i & \text{if } k_i = l_i, \\ X_i' & \text{if } k_i \neq l_i. \end{cases} \tag{12}$$

This assumption implies that the time order of the replicates is immaterial.

A5. For $i = 1, \ldots, m$, let the integer n_i^* denote $\max_{j=1}^{N} n_{ij}$. This n_i^* is free of N and provides an upper bound on the number of replications from the ith measurement method.

A6. Let $\mathbf{r}$ be the m-vector $(r_1, \ldots, r_m)$ and $p_N(\mathbf{r})$ denote the proportion of experimental units for whom the number of replications from the measurement methods $1, \ldots, m$ are $r_1, \ldots, r_m$, respectively. Here $r_i = 1, \ldots, n_i^*$, $i = 1, \ldots, m$. There exists a $p^*(\mathbf{r})$ such that as $N \to \infty$, $p_N(\mathbf{r}) \to p^*(\mathbf{r})$ for each $\mathbf{r}$. These proportions satisfy $\sum_{\mathbf{r}} p_N(\mathbf{r}) = 1 = \sum_{\mathbf{r}} p^*(\mathbf{r})$, where the symbol "$\sum_{\mathbf{r}}$" represents "$\sum_{r_1=1}^{n_1^*} \cdots \sum_{r_m=1}^{n_m^*}$."

A7. There exists a finite limit $w^*(\mathbf{r})$ such that as $N \to \infty$, $Nw(N, \mathbf{r}) \to w^*(\mathbf{r})$ for each $\mathbf{r}$, and w^*

satisfies $\sum_{\mathbf{r}} \{\prod_{i=1}^{m} r_i\} p^*(\mathbf{r}) w^*(\mathbf{r}) = 1$.

5.1   Limit distribution of $\hat{\boldsymbol{\theta}}$

We now show that the limit distribution of the plug-in estimator $\hat{\boldsymbol{\theta}} = \mathbf{h}(\hat{F})$, as $N \to \infty$, is p-variate

normal. This result is derived in two steps. The first step is to show that the stochastic process

$\{\hat{F}(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ approaches a Gaussian process in the limit. To this end, let $\mathcal{D} = \{aH_1 + bH_2 :$

$a, b \in \mathbb{R}; H_1, H_2 \in \mathcal{F}\}$ denote the linear space generated by the cdfs in the class $\mathcal{F}$ of m-variate

cdfs on $\mathcal{X}$. The space $\mathcal{D}$ is equipped with the sup norm, $\|H\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} |H(\mathbf{x})|$, $H \in \mathcal{D}$.

Lemma 1: Suppose that the assumptions A1-A7 hold. Then as $N \to \infty$, $N^{1/2}(\hat{F} - F)$ converges

in distribution to a zero-mean Gaussian process in $\mathcal{D}$.

Proof: For $\mathbf{x} \in \mathcal{X}$, write

$$N^{1/2}\{\hat{F}(\mathbf{x}) - F(\mathbf{x})\} = \sum_{\mathbf{r}} \{Nw(N, \mathbf{r})\} Z_{\mathbf{r}}^N(\mathbf{x}),$$

where $Z_{\mathbf{r}}^N(\mathbf{x}) = N^{-1/2} \sum_{\{j:\mathbf{n}_j=\mathbf{r}\}} \sum_{k_1=1}^{r_1} \cdots \sum_{k_m=1}^{r_m} \{I(X_{1jk_1} \leq x_1, \ldots, X_{mjk_m} \leq x_m) - F(\mathbf{x})\}$. By

proceeding as in the proof of the first part of Lemma A.1 in Olsson and Rootzén (1996), it can

be shown that for each $\mathbf{r}$ as $N \to \infty$, $Z_{\mathbf{r}}^N$ converges in distribution to $Z_{\mathbf{r}}$, which represents an

independent, tight, zero-mean Gaussian process in $\mathcal{D}$. Further from A7, $Nw(N, \mathbf{r}) \to w^*(\mathbf{r})$. Now

an application of Whitt (1980, Theorem 4.1) and the continuous mapping theorem (van der Vaart

1998, Theorem 18.11) shows that $N^{1/2}(\hat{F} - F)$ converges in distribution to $\sum_{\mathbf{r}} w^*(\mathbf{r}) Z_{\mathbf{r}}$, which is

a Gaussian process in $\mathcal{D}$ with mean zero. $\square$

This result generalizes Theorem 3.1 of Olsson and Rootzén (1996) concerning a univariate cdf

to a multivariate cdf for the case when the weights in $\hat{F}$ are free of $\mathbf{x}$. For the second step in

the derivation of the limit distribution of $\hat{\boldsymbol{\theta}}$, it is assumed that this functional is differentiable

in an appropriate sense, and the result follows from the functional delta method (van der Vaart,

Chapter 20). In particular, we assume that:

A8. For each $(u,v) \in \mathcal{S}$, the functional $h_{uv} : \mathcal{F} \subseteq \mathcal{D} \mapsto \mathbb{R}$ is *Hadamard differentiable* (van der

Vaart 1998, Chapter 20) at $F \in \mathcal{F}$ tangentially to $\mathcal{D}_0 \subseteq D$, i.e., there exists a continuous

linear map $h'_{uv,F} : \mathcal{D}_0 \mapsto \mathbb{R}$ such that for any real sequence $t \to 0$, and $\{H, H_t\} \in \mathcal{D}_0$ satisfying $H_t \to H$ and $F + tH_t \in \mathcal{F}$, we have

$$\lim_{t \to 0} \frac{h_{uv}(F + tH_t) - h_{uv}(F)}{t} = h'_{uv,F}(H). \tag{13}$$

The functional $h'_{uv,F}(H)$ is called the Hadamard derivative of h_{uv} at F in the direction H. Assume also that the map $h'_{uv,F} : \mathcal{D} \mapsto \mathbb{R}$ is defined and is continuous on entire $\mathcal{D}$.

Under the assumption A8, the influence function of θ_{uv} can be written as $L_{uv}(x_u, x_v) = h'_{uv,F}(\delta_{\mathbf{x}} - F)$, and the derivative $h'_{uv,F}(H) = \int_{\mathcal{X}} L_{uv}(\mathbf{x}, F)dH(\mathbf{x})$ (Fernholz 1983, Sections 2.2 and 4.4). This assumption also implies that the functional $\mathbf{h} : \mathcal{F} \subseteq \mathcal{D} \mapsto \mathbb{R}^p$ is Hadamard differentiable at $F \in \mathcal{F}$ tangentially to $\mathcal{D}_0$, with derivative $\mathbf{h}'_F : \mathcal{D}_0 \mapsto \mathbb{R}^p$, which is a p-vector with components $h'_{uv,F}$, $(u, v) \in \mathcal{S}$. Moreover, as a map, $\mathbf{h}'_F : \mathcal{D} \mapsto \mathbb{R}^p$ is defined and is continuous on all of $\mathcal{D}$. It also follows that $\mathbf{h}'_F(H) = \int_{\mathcal{X}} \mathbf{L}(\mathbf{x}, F)dH(\mathbf{x})$, where $\mathbf{L}(\mathbf{x}, F)$ is the influence function of $\boldsymbol{\theta}$. This function is a p-vector with components $L_{uv}(x_u, x_v)$, $(u, v) \in \mathcal{S}$, and satisfies $\int_{\mathcal{X}} \mathbf{L}(\mathbf{x}, F)dF(\mathbf{x}) = \mathbf{0}$.

We are now ready to state the asymptotic normality result. In this result, the diagonal elements of Σ consist of σ_{uv}^2, $(u, v) \in \mathcal{S}$, where σ_{uv}^2 represents the variance of the limit distribution of $N^{1/2}(\hat{\theta}_{uv} - \theta_{uv})$. Further, the off-diagonal elements of Σ consist of $\sigma_{uv,st}$, $(u, v) \neq (s, t) \in \mathcal{S}$, where $\sigma_{uv,st}$ represents the covariance of the joint limit distribution of $N^{1/2}(\hat{\theta}_{uv} - \theta_{uv})$ and $N^{1/2}(\hat{\theta}_{st} - \theta_{st})$.

Theorem 1: Suppose that the assumptions A1-A8 hold. Then as $N \to \infty$, $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \equiv N^{1/2}\{\mathbf{h}(\hat{F}) - \mathbf{h}(F)\}$ converges in distribution to a $\mathcal{N}_p(0, \Sigma)$ distribution, provided that Σ is finite and non-singular. Assume additionally that Σ can be obtained as the limiting covariance matrix of $N^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$. Then the elements of Σ are:

$$\sigma_{uv}^2 = \sum_{\mathbf{r}} w^{*2}(\mathbf{r}) \, p^*(\mathbf{r}) \frac{\prod_{i=1}^{m} r_i^2}{r_u r_v} \left[E\{L_{uv}^2(X_u, X_v)\} \right.$$

$$+ (r_u - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X'_u, X_v)\} + (r_v - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X_u, X'_v)\}$$

$$\left. + (r_u - 1)(r_v - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X'_u, X'_v)\} \right],$$

$$\sigma_{uv,st} = \sum_{\mathbf{r}} w^{*2}(\mathbf{r}) \, p^*(\mathbf{r}) \prod_{i=1}^{m} r_i^2 \times$$

17

$$
\begin{cases}
\frac{1}{r_u}\big[E\{L_{uv}(X_u,X_v)L_{ut}(X_u,X_t)\}+(r_u-1)E\{L_{uv}(X_u,X_v)L_{ut}(X'_u,X_t)\}\big], & s=u,t\neq v, \\[4pt]
\frac{1}{r_v}\big[E\{L_{uv}(X_u,X_v)L_{vt}(X_v,X_t)\}+(r_v-1)E\{L_{uv}(X_u,X_v)L_{vt}(X'_v,X_t)\}\big], & s=v,t\neq u, \\[4pt]
\frac{1}{r_v}\big[E\{L_{uv}(X_u,X_v)L_{sv}(X_s,X_v)\}+(r_v-1)E\{L_{uv}(X_u,X_v)L_{sv}(X_s,X'_v)\}\big], & s\neq u,t=v, \\[4pt]
\frac{1}{r_u}\big[E\{L_{uv}(X_u,X_v)L_{su}(X_s,X_u)\}+(r_u-1)E\{L_{uv}(X_u,X_v)L_{su}(X_s,X'_u)\}\big], & s\neq v,t=u, \\[4pt]
E\{L_{uv}(X_u,X_v)L_{st}(X_s,X_t)\}, & s,t\neq u,v.
\end{cases} \qquad (14)
$$

Proof: The asymptotic normality of $N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ follows from Lemma 1 and an application of the functional delta method (van der Vaart 1998, Theorem 20.8). We now focus on obtaining Σ. Since the derivative map in assumption A8 exists and is continuous on entire $\mathcal{D}$, the second part of the Theorem 20.8 cited above, implies that $N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$ and $\mathbf{h}'_F\{N^{1/2}(\hat{F}-F)\}$ have the same limit distribution. It follows that $\Sigma=\lim_{N\to\infty} var[\mathbf{h}'_F\{N^{1/2}(\hat{F}-F)\}]$ as Σ is assumed to be the limiting covariance matrix of $N^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})$. Next, using (3), we can write

$$
\begin{aligned}
\mathbf{h}'_F\{N^{1/2}(\hat{F}-F)\} &= \int_{\mathcal{X}} \mathbf{L}(\mathbf{x},F)\,d\{N^{1/2}(\hat{F}-F)(\mathbf{x})\} = N^{1/2}\int_{\mathcal{X}}\mathbf{L}(\mathbf{x},F)d\hat{F}(\mathbf{x}) \\
&= N^{1/2}\sum_{j=1}^{N} w(N,\mathbf{n}_j)\sum_{k_1=1}^{n_{1j}}\cdots\sum_{k_m=1}^{n_{mj}}\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F) \\
&= N^{1/2}\sum_{\mathbf{r}} w(N,\mathbf{r})\sum_{\{j:\mathbf{n}_j=\mathbf{r}\}}\sum_{k_1=1}^{r_1}\cdots\sum_{k_m=1}^{r_m}\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F).
\end{aligned}
$$

Now upon computing variance of both sides and noticing that, due to the assumption A4, the variance of $\sum_{k_1=1}^{r_1}\cdots\sum_{k_m=1}^{r_m}\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F)$ is free of j, we get

$$
\begin{aligned}
\Sigma &= \lim_{N\to\infty}\sum_{\mathbf{r}}\{N^2 w^2(N,\mathbf{r})\}\,p_N(\mathbf{r})\,var\left\{\sum_{k_1=1}^{r_1}\cdots\sum_{k_m=1}^{r_m}\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F)\right\} \\
&= \sum_{\mathbf{r}} w^{*2}(\mathbf{r})\,p^*(\mathbf{r})\,var\left\{\sum_{k_1=1}^{r_1}\cdots\sum_{k_m=1}^{r_m}\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F)\right\}. \qquad (15)
\end{aligned}
$$

To find σ^2_{uv}, consider $L_{uv}(X_{ujk_u},X_{vjk_v})$, the element of vector $\mathbf{L}((X_{1jk_1},\ldots,X_{mjk_m}),F)$ that corresponds to $\hat{\theta}_{uv}$. Upon using (12) and the fact that $L_{uv}(X_{ujk_u},X_{vjk_v})$ has mean zero,

$$
var\left\{\sum_{k_1=1}^{r_1}\cdots\sum_{k_m=1}^{r_m}L_{uv}(X_{ujk_u},X_{vjk_v})\right\}=\frac{\prod_{i=1}^{m}r_i^2}{r_u^2 r_v^2}\sum_{k_u=1}^{r_u}\sum_{k_v=1}^{r_v}\sum_{l_u=1}^{r_u}\sum_{l_v=1}^{r_v}E\{L_{uv}(X_u,X_v)L_{uv}(Y_u,Y_v)\}
$$

18

$$= \frac{\prod_{i=1}^{m} r_i^2}{r_u r_v} \Big[E\{L_{uv}^2(X_u, X_v)\} + (r_u - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X_u', X_v)\}$$

$$+ (r_v - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X_u, X_v')\} + (r_u - 1)(r_v - 1)E\{L_{uv}(X_u, X_v)L_{uv}(X_u', X_v')\} \Big].$$

The expression for σ_{uv}^2 now follows by substituting the above expression in (15). Next, for $\sigma_{uv,st}$, consider two elements of $\mathbf{L}((X_{1jk_1}, \ldots, X_{mjk_m}), F)$, namely, $L_{uv}(X_{ujk_u}, X_{vjk_v})$ and $L_{st}(X_{sjk_s}, X_{tjk_t})$, that correspond to $\hat{\theta}_{uv}$ and $\hat{\theta}_{st}$, respectively. Further,

$$cov \left\{ \sum_{k_1=1}^{r_1} \cdots \sum_{k_m=1}^{r_m} L_{uv}(X_{ujk_u}, X_{vjk_v}), \sum_{k_1=1}^{r_1} \cdots \sum_{k_m=1}^{r_m} L_{st}(X_{sjk_s}, X_{tjk_t}) \right\} = \frac{\prod_{i=1}^{m} r_i^2}{r_u r_v r_s r_t}$$

$$\times \sum_{k_u=1}^{r_u} \sum_{k_v=1}^{r_v} \sum_{l_s=1}^{r_s} \sum_{l_t=1}^{r_t} E\{L_{uv}(X_{ujk_u}, X_{vjk_v})L_{st}(X_{sjl_s}, X_{tjl_t})\}. \tag{16}$$

Now there are five possibilities depending on whether or not there is a measurement method that is involved in both $\hat{\theta}_{uv}$ and $\hat{\theta}_{st}$. They are: $(s = u, t \neq v)$, $(s = v, t \neq u)$, $(s \neq u, t = v)$, $(s \neq v, t = u)$ and $(s, t \neq u, v)$. When $(s = u, t \neq v)$, the quadruple sum on the right in (16) can be written as $r_u r_v r_t \big[E\{L_{uv}(X_u, X_v)L_{ut}(X_u, X_t)\} + (r_u - 1)E\{L_{uv}(X_u, X_v)L_{ut}(X_u', X_t)\} \big]$. Similar expressions for the sum in (16) can be derived for the remaining four cases as well. The result then follows upon substitution of these expressions in (16) and (15). $\qquad \square$

It may be noted that when the measurements are unreplicated, i.e., $n_{ij} = 1$ for all (i, j), σ_{uv}^2 and $\sigma_{uv,st}$ defined in (14) reduce to $E\{L_{uv}^2(X_u, X_v)\}$ and $E\{L_{uv}(X_u, X_v)L_{st}(X_s, X_t)\}$, respectively.

5.2 A consistent estimator $\hat{\Sigma}$ for Σ

Recall that $\hat{L}_{uv}(x_u, x_v) \equiv L_{uv}(\mathbf{x}, \hat{F})$ denotes the empirical counterpart of the influence function $L_{uv}(x_u, x_v) \equiv L_{uv}(\mathbf{x}, F)$, $(u, v) \in \mathcal{S}$. To obtain $\hat{\Sigma}$, we simply replace $w^*(\mathbf{r})$ in Σ, given by (14), with $Nw(N, \mathbf{r})$, $p^*(\mathbf{r})$ with $p_N(\mathbf{r})$, and the population moments of $L_{uv}(X_u, X_v)$ with the sample moments of $\hat{L}_{uv}(X_u, X_v)$. Note that $p_N(\mathbf{r}) = 1$ when $\mathbf{r} = \mathbf{n}_j$, the vector of the observed number of replications, otherwise $p_N(\mathbf{r}) = 0$. We next describe the estimators of the moments of $L_{uv}(X_u, X_v)$. The four moments in σ_{uv}^2, namely, $E\{L_{uv}^2(X_u, X_v)\}$, $E\{L_{uv}(X_u, X_v)L_{uv}(X_u', X_v)\}$, $E\{L_{uv}(X_u, X_v)L_{uv}(X_u, X_v')\}$ and $E\{L_{uv}(X_u, X_v)L_{uv}(X_u', X_v')\}$, can be respectively estimated by:

$$\frac{1}{N} \sum_{j=1}^{N} \frac{1}{n_{u_j} n_{v_j}} \sum_{k_u=1}^{n_{u_j}} \sum_{k_v=1}^{n_{v_j}} \hat{L}_{uv}^2(X_{ujk_u}, X_{vjk_v}),$$

19

$$\frac{1}{\#\{j:n_{u_j}>1\}}\sum_{j=1}^{\#\{j:n_{u_j}>1\}}\frac{1}{n_{u_j}n_{v_j}(n_{u_j}-1)}\sum_{k_u=1}^{n_{u_j}}\sum_{k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjk_v})\sum_{l_u\neq k_u=1}^{n_{u_j}}\hat{L}_{uv}(X_{ujl_u},X_{vjk_v}),$$

$$\frac{1}{\#\{j:n_{v_j}>1\}}\sum_{j=1}^{\#\{j:n_{v_j}>1\}}\frac{1}{n_{u_j}n_{v_j}(n_{v_j}-1)}\sum_{k_u=1}^{n_{u_j}}\sum_{k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjk_v})\sum_{l_v\neq k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjl_v}),$$

$$\frac{1}{\#\{j:n_{u_j}>1,n_{v_j}>1\}}\sum_{j=1}^{\#\{j:n_{u_j}>1,n_{v_j}>1\}}\frac{1}{n_{u_j}n_{v_j}(n_{u_j}-1)(n_{v_j}-1)}\sum_{k_u=1}^{n_{u_j}}\sum_{k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjk_v})$$

$$\times\sum_{l_u\neq k_u=1}^{n_{u_j}}\sum_{l_v\neq k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujl_u},X_{vjl_v}).$$

The moment of the form $E\{L_{uv}(X_u,X_v)L_{st}(X_s,X_t)\}$ in $\sigma_{uv,st}$, can be estimated as:

$$\frac{1}{N}\sum_{j=1}^{N}\frac{1}{n_{u_j}n_{v_j}n_{s_j}n_{t_j}}\sum_{k_u=1}^{n_{u_j}}\sum_{k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjk_v})\sum_{k_s=1}^{n_{s_j}}\sum_{k_t=1}^{n_{t_j}}\hat{L}_{st}(X_{sjk_s},X_{tjk_t}),$$

where if $s=u$ (or $s=v$), n_{s_j} is removed from the denominator and the sum over k_s is restricted to $k_s=k_u$ (or $k_s=k_v$); and a similar modification is made if $t=u$ or $t=v$. Further, $E\{L_{uv}(X_u,X_v)L_{ut}(X'_u,X_t)\}$ in $\sigma_{uv,st}$ can be estimated as

$$\frac{1}{\#\{j:n_{u_j}>1\}}\sum_{j=1}^{\#\{j:n_{u_j}>1\}}\frac{1}{n_{u_j}n_{v_j}(n_{u_j}-1)n_{t_j}}\sum_{k_u=1}^{n_{u_j}}\sum_{k_v=1}^{n_{v_j}}\hat{L}_{uv}(X_{ujk_u},X_{vjk_v})$$

$$\times\sum_{l_u\neq k_u=1}^{n_{u_j}}\sum_{k_t=1}^{n_{t_j}}\hat{L}_{ut}(X_{ujl_u},X_{tjk_t}),$$

and similar estimators can be constructed for the remaining three moments in $\sigma_{uv,st}$, namely, $E\{L_{uv}(X_u,X_v)L_{vt}(X'_v,X_t)\}$, $E\{L_{uv}(X_u,X_v)L_{sv}(X_s,X'_v)\}$ and $E\{L_{uv}(X_u,X_v)L_{su}(X_s,X'_u)\}$.

5.3   Validity of the confidence intervals for θ

Our next result establishes the validity of the confidence intervals given in (6).

Theorem 2: Under the assumptions of Theorem 1, the simultaneous coverage probability of each of the three sets of confidence intervals in (6) converges to $(1-\alpha)$ as $N\to\infty$.

Proof: The result is proved only for the upper bounds as similar arguments can be used for the lower bounds and the two-sided intervals. The coverage probability of the upper bounds can be written as $P\left(\min_{(u,v)\in\mathcal{S}}N^{1/2}(\hat{\theta}_{uv}-\theta_{uv})/\hat{\sigma}_{uv}+\hat{c}_{1-\alpha,p}\geq 0\right)$. Now an application of Theorem 1,

Slutsky's theorem (Lehmann 1998, Theorem 2.3.3) and the continuous mapping theorem (van der Vaart 1998, Theorem 18.11) implies that $\min_{(u,v)\in\mathcal{S}} N^{1/2}(\hat{\theta}_{uv} - \theta_{uv})/\hat{\sigma}_{uv}$ converges in distribution to $\min_{(u,v)\in\mathcal{S}} Z_{uv}$. Moreover, $\hat{c}_{1-\alpha,p}$ converges in probability to $c_{1-\alpha,p}$. The result now follows from another application of Slutsky's theorem and noting that $\min_{(u,v)\in\mathcal{S}} Z_{uv}$ and $-\max_{(u,v)\in\mathcal{S}} Z_{uv}$ have the same distribution. □

5.4 Results for the four agreement measures

We now verify the assumption of Hadamard differentiability (A8) for the four measures studied in Section 2.4 and confirm the expressions for their influence functions given in (8).

Theorem 3:

(a) Suppose that the measurements from the m measurement methods lie in $[-M, M]$, for some finite positive M. Then for $(u, v) \in \mathcal{S}$, the assumption A8 holds for $MSD_{uv}(F)$ and $CCC_{uv}(F)$.

(b) For $(u, v) \in \mathcal{S}$, the assumption A8 holds for $CP_{uv}(F)$.

(c) Suppose that for every $(u, v) \in \mathcal{S}$, the cdf G_{uv} is differentiable at TDI_{uv} with positive derivative $g_{uv}(TDI_{uv})$. Then the assumption A8 holds for $TDI_{uv}(F)$.

(d) The expressions given in (8) hold for the influence functions of MSD_{uv}, CCC_{uv}, CP_{uv} and TDI_{uv}, $(u, v) \in \mathcal{S}$.

Proof: (a) Under the assumption, the support of the distribution of $\mathbf{X}$, i.e., $\mathcal{X} = [-M, M]^m$, is finite. This implies that the first and second order moments of $\mathbf{X}$, which are linear functionals of $F \in \mathcal{D}$, are also bounded maps from $\mathcal{D} \mapsto \mathbb{R}$. Therefore, these moments are continuous functionals. Now, as in Guo and Manatunga (2007, Lemma A.1), it can be seen that these moments are Hadamard differentiable at F with the derivative evaluated at $H \in \mathcal{D}$ equal to the same moments under H. It follows that the Hadamard derivative of $MSD_{uv}(F)$ at H is $\int_{\mathcal{X}} (x_u - x_v)^2 \, dH(\mathbf{x})$. Moreover, since $CCC_{uv}(F)$ is also a function of these moments, the chain rule for Hadamard

differentiability (van der Vaart 1998, Theorem 20.9) implies that the assumption A8 holds for $CCC_{uv}(F)$ as well.

The finiteness of support is not needed for (b) and (c). For (b), $CP_{uv}(F)$ is the expected value of an indicator (and hence a bounded) random variable. So, the arguments similar to above show that $CP_{uv}(F)$ is Hadamard differentiable at F with derivative at $H \in \mathcal{D}$ equal to $\int_{\mathcal{X}} I(|x_u - x_v| \leq \delta)\, dH(\mathbf{x})$. Thus, the assumption A8 holds for $CP_{uv}(F)$. For (c), since $TDI_{uv}(F)$ is a quantile, it follows from van der Vaart (1998, Lemma 21.3) that $TDI_{uv}(F)$ is Hadamard differentiable at F tangentially to the set of functions $H \in \mathcal{D}$ for which the function $H_{uv}(t) = \int_{\mathcal{X}} I(|x_u - x_v| \leq t) dH(\mathbf{x})$ is continuous at $t = TDI_{uv}$, with derivative at H equal to $-H_{uv}(TDI_{uv})/g_{uv}(TDI_{uv})$. Moreover, as a map from $\mathcal{D} \mapsto \mathbb{R}$, $-H_{uv}(TDI_{uv})/g_{uv}(TDI_{uv})$ exists and is continuous for all $H \in \mathcal{D}$. Thus, the assumption A8 holds for TDI_{uv} as well.

Finally, the expressions for the influence functions of these measures follow from evaluating their Hadamard derivatives at $H = \delta_{\mathbf{x}} - F$. $\qquad\square$

The next result establishes the validity of the confidence intervals for TDI_{uv}, given in (9).

Theorem 4: Let $0 < \nu_1 < \nu_2 < 1$ and ν be an interior point of $[\nu_1, \nu_2]$. Assume that for each $(u, v) \in \mathcal{S}$, G_{uv} is continuously differentiable on the interval $[a_{uv}, b_{uv}] = [G_{uv}^{-1}(\nu_1) - \epsilon, G_{uv}^{-1}(\nu_2) + \epsilon]$ for some $\epsilon > 0$, with strictly positive derivative g_{uv}. Then under the assumptions of Theorem 1, the simultaneous coverage probability of each set of intervals in (9) converges to $(1-\alpha)$ as $N \to \infty$.

Proof: We will only consider the upper bounds as the proof for the two-sided intervals is similar. First, write the coverage probability of the upper bounds as

$$P\left[TDI_{uv} \leq \hat{G}^{-1}(\nu + \hat{c}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2}),\ \text{for all } (u,v) \in \mathcal{S}\right]$$
$$= P\left[T_{uv} \geq g_{uv}(TDI_{uv})N^{1/2}\{\widehat{TDI}_{uv} - \hat{G}^{-1}(\nu + \hat{c}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2})\}/\hat{\sigma}_{uv},\ \text{for all } (u,v) \in \mathcal{S}\right],$$

where $T_{uv} = g_{uv}(TDI_{uv})N^{1/2}(\widehat{TDI}_{uv} - TDI_{uv})/\hat{\sigma}_{uv}$. Now, from an application of Theorem 3(c), Theorem 1 with $\boldsymbol{\theta}$ as the p-vector of TDI_{uv}, $(u,v) \in \mathcal{S}$, and the Slutsky's theorem, it follows that the asymptotic joint distribution of the T's is same as the p-variate normal distribution of the Z's defined for the intervals in (6). The result will now follow as in Theorem 2 provided one can show that $g_{uv}(TDI_{uv})N^{1/2}\{\widehat{TDI}_{uv} - \hat{G}^{-1}(\nu + \hat{c}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2})\}/\hat{\sigma}_{uv}$ converges in probability to $-c_{1-\alpha,p}$.

For this, one can proceed along the lines of van der Vaart (1998, Lemma 21.7) to show

that the difference between $N^{1/2}\{TDI_{uv} - G^{-1}(\nu + \hat{c}_{1-\alpha,p}\,\hat{\sigma}_{uv}/N^{1/2})\}$ and $N^{1/2}\{TDI_{uv} - G^{-1}(\nu +$

$c_{1-\alpha,p}\,\sigma_{uv}/N^{1/2})\}$ converges in probability to zero. Consequently, the first term converges in prob-

ability to $-c_{1-\alpha,p}\,\sigma_{uv}/g_{uv}(TDI_{uv})$ as it is the limit of the second term. The desired result now

holds because $\hat{\sigma}_{uv}$ is consistent for σ_{uv}, for each $(u,v) \in \mathcal{S}$. $\qquad\qquad\square$

6   Discussion

This article presents a unified nonparametric approach for multiple comparisons using a measure of

agreement. It assumes that the repeated measurements from multiple measurement methods are

unpaired. The proposed methodology can also handle the case when the repeated measurements

are paired by simply using the observed m-tuples of measurements in the estimation instead of

using the all possible m-tuples of unpaired measurements. Furthermore, when more than two

measurement methods are compared, rather than performing multiple comparisons as we do in

this article, many authors (e.g., Barnhart, Haber and Song 2002, and Lin et al. 2007) perform

inference on a single index that measures the overall level of agreement among all methods. Since

this overall measure is also a functional of F, the proposed nonparametric methodology can deal

with this situation as well.

This article focuses on evaluation of agreement *between* measurement methods. But when the

measurements are repeated, one may be additionally interested in the evaluation of agreement of

a method with itself as the extent of this within-method agreement serves as a baseline for the

evaluation of between-method agreement (Bland and Altman 1999). A within-method agreement

measure is a functional of the joint cdf of two replicate measurements from a method on a randomly

selected experimental unit from the population. As in (3), this cdf can be estimated using a

weighted empirical cdf and the proposed methodology can be adapted to handle this situation.

In this article, we assume homogeneity, i.e., the measurements on different experimental units

are identically distributed. This assumption is violated, e.g., when the distribution of measure-

ments depends on a covariate. The proposed methodology can be generalized to deal with cate-

23

gorical covariates by estimating F separately within each category of the covariate. The extension

to deal with continuous covariates is a topic of future research.

References

Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. *Journal of the Royal Statistical Society, Series B* **65**, 367–389.

Barnhart, H. X. and Williamson, J. M. (2001). Modeling concordance correlation via GEE to evaluate reproducibiltiy. *Biometrics* **57**, 931–940.

Barnhart, H. X., Haber, M. J. and Lin, L. I. (2007). An overview on assessing agreement with continuous measurement. *Journal of Biopharmaceutical Statistics* **17**, 529–569.

Barnhart, H. X., Haber, M. J. and Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics* **58**, 1020–1027.

Barnhart, H. X., Song, J. and Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine* **24**, 1371–1384.

Bland, J. M. and Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **i**, 307–310.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* **8**, 135–160.

Carrasco, J. L. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics* **59**, 849–858.

Carstensen, B., Simpson, J. and Gurrin, L. C. (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* **4**, issue 1, article 16, available at `http://www.bepress.com/ijb/vol4/iss1/16`.

Chinchilli, V. M., Martel, J. K., Kumanyika, S. and Lloyd, T. (1996). A weighted concordance correlation coefficient for repeated measurement designs. *Biometrics* **52**, 341–353.

Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference* **138**, 1102–1115.

Choudhary, P. K. and Nagaraja, H. N. (2007). Tests for assessment of agreement using probability criteria. *Journal of Statistical Planning and Inference* **137**, 279–290.

Choudhary, P. K. and Yin, K. (2010). Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research.* To appear.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application.* Cambridge University Press, New York.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals.* Springer-Verlag, New York.

Guo, Y. and Manatunga, A. K. (2007). Nonparametric estimation of the concordance correlation coefficient under univariate censoring. *Biometrics* **83**, 164–172.

Hawkins, D. M. (2002). Diagnostics for conformity of paired quantitative measurements. *Statistics in Medicine* **21**, 1913–1935.

Hothorn, T., Bretz, F. and Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal* **50**, 346–363.

King, T. S. and Chinchilli, V. M. (2001a). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine* **20**, 2131–2147.

King, T. S. and Chinchilli, V. M. (2001b). Robust estimators of the concordance correlation coefficient. *Journal of Biopharmaceutical Statistics* **11**, 83–105.

King, T. S., Chinchilli, V. M. and Carrasco, J. L. (2007a). A repeated measures concordance correlation coefficient. *Statistics in Medicine* **26**, 3095–3113.

King, T. S., Chinchilli, V. M., Wang, K.-L. and Carrasco, J. L. (2007b). A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics* **17**, 653–672.

Lehmann, E. L. (1998). *Elements of Large Sample Theory*. Springer-Verlag, New York.

Lin, L., Hedayat, A. S. and Wu, W. (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics* **17**, 629–652.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268. Corrections: 2000, 56:324-325.

Lin, L. I. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine* **19**, 255–270.

Lin, L. I., Hedayat, A. S., Sinha, B. and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association* **97**, 257–270.

Olsson, J. and Rootzén, H. (1996). Quantile estimation from repeated measurements. *Journal of the American Statistical Association* **91**, 1560–1565.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and the R Core team (2009). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-92.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.

Whitt, W. (1980). Some useful functions for functional limit theorems. *Mathematics of Operations Research* **5**, 67–85.
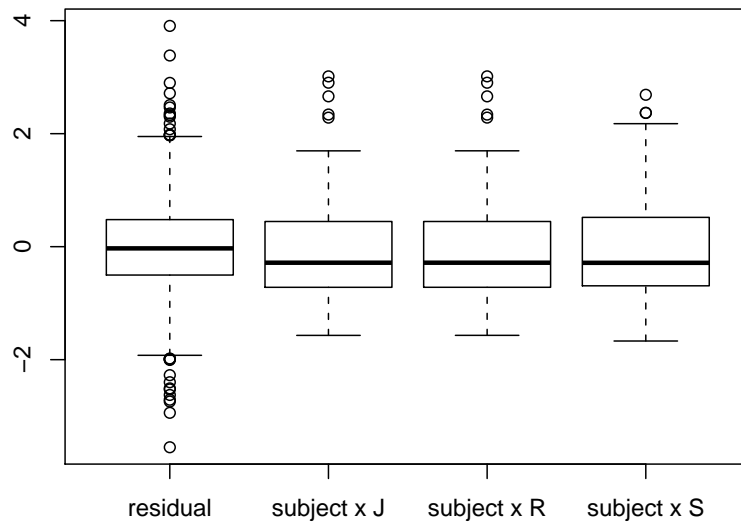
Figure 1: Box plots of standardized residuals and estimates of random subject-by-measurement method interaction effects when the mixed-effects model (10) is fitted to the blood pressure data.

		Nonparametric			Model-based		
		JR	JS	RS	JR	JS	RS
CCC	estimate	0.97	0.70	0.70	0.97	0.70	0.70
	std. error	0.01	0.08	0.08	0.01	0.05	0.05
	95% lower bound	0.96	0.52	0.52	0.96	0.59	0.59
TDI	estimate	12	34	35	12.9	42.6	42.6
	std. error	-	-	-	0.4	2.8	2.7
	95% upper bound	14	54	53	13.8	48.9	48.7

Table 1: Summary of results for the blood pressure data. The standard errors for TDI estimates are omitted as they are not needed for bounds computed using (9) to avoid density estimation.

	$N = 30$				$N = 60$				$N = 100$			
	Normal		Skew-t		Normal		Skew-t		Normal		Skew-t	
n	CCC	TDI	CCC	TDI	CCC	TDI	CCC	TDI	CCC	TDI	CCC	TDI
1	93.2	89.8	91.2	89.5	94.3	96.2	94.3	96.7	94.4	98.0	95.1	98.2
2	95.0	94.3	93.4	94.4	94.9	95.0	95.2	95.1	95.3	95.4	95.2	94.8
3	94.4	92.9	93.1	93.2	94.6	94.1	95.2	94.9	95.4	95.3	96.6	95.0
4	94.4	93.2	94.1	91.6	95.6	93.6	94.8	93.4	95.0	94.2	95.8	93.6

Table 2: Estimated coverage probabilities (%) of asymptotic 95% simultaneous confidence bounds.

	Normal				Skew-t			
	CCC		TDI		CCC		TDI	
	Low	High	Low	High	Low	High	Low	High
Balanced designs (nonparametric vs. ML)								
$n = 1$	1.00	1.00	1.59	1.59	1.00	1.00	0.35	0.33
$n = 2$	1.00	1.00	1.25	1.32	1.00	1.00	0.44	0.36
$n = 3$	1.00	1.00	1.17	1.23	1.00	1.00	0.43	0.28
$n = 4$	1.00	1.00	1.15	1.14	1.00	1.00	0.39	0.26
Unbalanced design (nonparametric vs. ML)								
$w = w_1$	1.01	1.10	1.38	1.98	1.00	1.02	0.44	0.44
$w = w_2$	1.49	1.46	1.72	1.29	1.46	1.43	0.59	0.41
Unbalanced design (w_2 vs. w_1)								
$N = 500$	1.48	1.33	1.25	0.65	1.46	1.40	1.34	0.93
$N = \infty$	1.49	1.31	1.25	0.68	1.56	1.46	1.37	0.91

Table 3: Approximate AREs of various pairs of estimators. "Low" and "High" refer to parameter combinations that respectively correspond to low agreement and high agreement between two measurement methods. "$N = \infty$" refers to ARE defined as the ratio of asymptotic variances.