Published in final edited form as:

Stat Med. 2011 December 20; 30(29): 3403–3415. doi:10.1002/sim.4366.

Simple Estimation of Hidden Correlation in Repeated Measures

Thuan Nguyen and Jiming Jiang

Oregon Health and Science University and University of California, Davis

Abstract

In medical and social studies, it is often desirable to assess the correlation between characteristics of interest that are not directly observable. In such cases, repeated measures are often available, but the correlation between the repeated measures is not the same as that between the true characteristics that are confounded with the measurement errors. The latter is called the hidden correlation. Previously, the problem has been treated by assuming prior knowledge about the measurement errors, or by using relatively complex statistical models, such as the mixed effects models, with no closed-form expression for the estimated hidden correlation. We propose a simple estimator of the hidden correlation that is very much like the Pearson correlation coefficient, with a closed-form expression, under assumptions much weaker than the mixed effects model. Simulation results show that the proposed simple estimator performs similarly as the restricted maximum likelihood (REML) estimator in mixed models, but is computationally much more efficient than REML. Simulation comparison with the Pearson correlation is also made. A real data example is considered.

Keywords

Correlation Coefficient; Hypothesis Testing; Repeated Measures

1 Introduction

In medical and social studies, it is often desirable to assess the correlation between characteristics of interest in the presence of repeated measures. For example, the repeated measures may be obtained under different circumstances, such as (1) by different observers (e.g., doctors, nurses, technicians); (2) at different occasions (e.g., times, visits); or (3) from different individuals sharing the same characteristics. If the characteristics of interest are directly observed, a well-known simple measure of the correlation between them is the Pearson correlation, also known as the correlation coefficient (e.g., [1]), computed by

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (1)$$

where (x_i, y_i) is the pair of observations from subject i , $i = 1, \dots, n$, $\bar{x} = n^{-1} \sum_{i=1}^n x_i$ and $\bar{y} = n^{-1} \sum_{i=1}^n y_i$. Here r is considered an estimator of the population correlation coefficient between x and y . However, in many cases, the variables of interest are not directly observable; instead, repeated, “indirect”, measures of these variables are available. It should be noted that the term “repeated measures” should be understood in a broader sense as any observations that involve errors so that the true characteristic of interest is not observed. We use a real-data example to illustrate.

Example

Here we consider a daily study conducted by Dr. Ryan Olson and his colleagues at Oregon Health & Science University to measure home care worker's exposure levels to demanding tasks, and also to understand how work exposure might relate to daily physical symptoms, wellness, and lifestyle behaviors. Twenty-three home care workers participated in the study. After undertaking a baseline demographic and psychosocial survey, the participants completed 10–14 days of daily self-monitoring using the safety task assessment tool developed by Dr. Olson's research group. It was noted that the participants might meet different clients on those days. The work task survey focuses on measuring time and frequency exposures to demanding tasks, and was completed by the worker during the natural breaks at work. The end-of-day survey focuses on recording physical, behavior, and psychosocial symptoms that could be affected by the work task exposures, and was completed in the evening between dinner and bedtime. In short, the daily data include the 24 variables recording the counts and times spent on various tasks. The correlations between these variables were of interest, for example, that between Daily Occupational Fatigue and Daily Negative Affect. However, these variables were recorded based on the workers' estimates, or memories, say, during the natural breaks. Such estimates might not be accurate. Besides, there is variation due to the occasion (i.e., day) that exists even if the workers report their daily activities accurately. In other words, the true characteristics of interest (e.g., Daily Occupational Fatigue Stress and Daily Negative Affect) were measured with combinations of different sources of errors.

Nevertheless, with the repeated measures, we may be able to recover the true correlation between the characteristics of interest. Let (x_i, y_i) denote the pair of unobserved characteristics of interest for subject i , and (x_{is}, y_{is}) , $s = 1, \dots, n_i$ the repeated measures of (x_i, y_i) . Imagine that x_{is} is a combination of the true value x_i plus a measurement error (that combines different sources of variations), that is, $x_{is} = x_i + \xi_{is}$, where ξ_{is} is the measurement error for x_i . Similarly, $y_{is} = y_i + \eta_{is}$ and η_{is} is the measurement error for y_i . For example, in the example above, x_i is the mean Daily Occupational Fatigue and y_i is the mean Daily Negative Affect, for worker i . Our interest is $\rho_h = \text{cor}(x_i, y_i)$, the population correlation coefficient for the pair of characteristics across all the workers (here the subscript h refers to "hidden correlation"; see below). However, this correlation is confounded with the measurement errors. For example, the latter often add noise to the "signals" (i.e., x_i and y_i), making one think that the signals are less correlated. In other words, the correlation between the signals is hidden, at least to some extent, by the noise. In fact, even if the measurement errors ξ_{is} and η_{is} are uncorrelated, the correlation between x_i and y_i is not the same as that between the repeated measures, x_{is} and y_{is} , but larger in absolute value than the latter, due to the variations of the measurement errors (see sections 2 and 4 for more details). It should be pointed out that the repeated measures do not create the hidden correlation problem; on the contrary, they provide a way to solve the problem. In fact, the hidden correlation can never be recovered without the repeated measures.

Also note that the problem considered here is different from the one regarding the intraclass correlation coefficient (ICC; e.g., [2]). The latter was originally proposed by R. A. Fisher [3] to assess consistency, or conformity, of measurements made by multiple observers measuring the same quantity. Also see [4]. The following example may help to highlight the difference. Each blood sample drawn from an autistic patient is examined by several medical laboratories to obtain measures of total cholesterol and 7-dehydrocholesterol. If one wants to know the consistency of the medical laboratories on their cholesterol measurements (over the population of all the patients), one has an ICC problem. On the other hand, if one wants to know the population correlation (across all the patients) between cholesterol and 7-dehydrocholesterol, one has a hidden correlation problem. Note that here the focus is the

two variables, and the (repeated) measures (by the medical laboratories) are the tool to help capture the correlation more accurately.

Some studies involve more than two variables (e.g., the Example above on home care workers). If one is interested in the correlations between these variables, by considering one pair of variables each time, these correlations can be estimated, provided that the above problem regarding the hidden correlation between x_i and y_i can be solved.

The hidden correlation problem has been studied in the literature. In most cases, it has been assumed that the measurement errors, ξ_{is} and η_{is} , are uncorrelated. For example, “case 3” of [5] (pp. 217–218) assumed that measurement errors are independent and normally distributed. However, the author did not provide a satisfactory solution to the problem, noting the “difficulty” due to the non-observability of the variables of interest. Following the same assumptions, [6] considered the problem as that of estimating the variance components, namely, the variances of x , y , ξ , η and ρ_h . The authors used the analysis of variance (ANOVA) method to estimate the variances, and then a formula simplified by the assumption, $c_{\xi\eta} = 0$ [see (3) below], to estimate ρ_h . [7] used the simplified connection (4) below with $c_{\xi\eta} = 0$, and assumed, in addition, that c_{xx} , c_{yy} , $c_{\xi\xi}$, $c_{\eta\eta}$ are obtained from “reproducibility studies”. Also see [8], [9] in regression error-in-variable settings. Unlike the above, [10] noted that it is sometimes unreasonable to assume that the measurement errors are uncorrelated, and the effects of ignoring the correlation. The authors proposed a method without assuming $c_{\xi\eta} = 0$, but they required that estimates of $c_{\xi\xi}$, $c_{\eta\eta}$ and $c_{\xi\eta}$ are “obtained elsewhere”; in other words, knowledge about the variances and correlation of the measurement errors is available. Also see [11], who considered the regression error-in-variable setting with $c_{\xi\eta} = 0$, but the focus was estimation of the regression coefficients.

A “modern” approach to estimation of the hidden correlation is to use a mixed effects model (e.g., [12]). For the most part, under a mixed effects model, the x_i and y_i are treated as random effects, whose correlation ρ_h is part of the so-called variance components. Thus, by estimating the variance components using, say, the restricted maximum likelihood (REML; e.g., [13]) method, one obtains an estimate of ρ_h . See section 4 for more detail. A procedure based on the mixed-effects-model approach is available in PROC MIXED in SAS (see [14]). However, in order to use this procedure to estimate the hidden correlation one needs to know the mixed effects models very well to understand which particular model should be used. This is often not the case to non-statistician users, not to mention that methodology background behind it. As many medical or social researchers are more familiar with the Pearson correlation, it might be more appealing to these researchers if an estimate of the hidden correlation can be computed in a way similar to the Pearson correlation, in particular, with a closed-form expression. Furthermore, the mixed-effects-model approach is based a highly parametric model involving the normality assumption, which may not be satisfied in practice. For example, in many medical or social studies the repeated measures are discrete (e.g., counts, integer-valued levels, as in the Example above on home care workers). Such data are obviously not normal, so the normality-based mixed effects model is questionable. Also, as noted by [10], it is impractical to assume $c_{\xi\eta} = 0$ in many cases. For example, if the repeated measures are done by different observers (e.g., doctors, nurses), there may be a correlation between the measurement errors, ξ_{is} and η_{is} , as they are due to the same technician. Therefore, we are not going to assume $c_{\xi\eta} = 0$, or any knowledge about the variances/covariance of the measurement errors, as in those previous work (with the exception of the mixed-effects-model approach), and yet still obtain a simple estimator of the hidden correlation as well as its standard error. These are some of the important features of our work.

A simple model is proposed in the next section and a closed-form estimator of the hidden correlation ρ_h is derived in a way very much like (1) under this model. The derivation shows, in particular, that blindly using the Pearson correlation, ignoring the difference between the true (hidden) correlation and the correlation between the observations, can lead to incorrect, or inaccurate, estimate of the true correlation. In section 3, we derive large-sample tests for the hidden correlation, which are often of interest in medical/social studies. Simulation studies are carried out in section 4 to evaluate the performance of the proposed estimator in terms of both estimation and testing. The simulation studies also include comparisons of our method with the Pearson correlation, and REML based on a linear mixed model. The simulation results support the theoretical finding that blindly using the Pearson correlation can be seriously misleading. Furthermore, the simulation results show that our method performs similarly to the REML method in terms of estimation and testing performance, especially when the sample size is relatively large, but (our method) is computationally much more efficient than REML. The Example above on the home care worker study is revisited in section 5. We highlight the main results in section 6. Discussions on potential use of bootstrap and other alternative methods are also offer in section 6. Technical details are deferred to Appendix.

2 A statistical model and derivation of estimator

Suppose that $x_{is}, y_{is}, i = 1, \dots, m; s = 1, \dots, n_i$ are observed, where $n_i > 1$, such that

$$x_{is} = x_i + \xi_{is}, y_{is} = y_i + \eta_{is}, \quad (2)$$

where i represents the subject so that (x_{is}, y_{is}) is the s^{th} repeated pair of measures from subject i , and n_i is the number of repeated measures from subject i , which is allowed to depend on i . Furthermore, x_i, y_i represent the characteristics of interest that are not observable, but do not depend on s , and ξ_{is}, η_{is} are the random fluctuations, or errors, due to the repeated measures. The following assumptions are made.

- A1. The vectors $(x_i, y_i), i = 1, \dots, m$ are i.i.d.
- A2. The vectors $(\xi_{is}, \eta_{is}), i = 1, \dots, m; s = 1, \dots, n_i$ are i.i.d. with mean 0.
- A3. For each $i, (x_i, y_i)$ and $(\xi_{is}, \eta_{is}), s = 1, \dots, n_i$ are uncorrelated.

Assumption A1 is the basis for the population correlation coefficient to be well-defined. Assumption A2 is reasonable if there is no bias against any subject or any repeat measures within the subject. For example, if the pairs (ξ_{is}, η_{is}) are viewed as random samples from the population of pairs of measurement errors for all the medical workers, then assumption A2 is expected to be satisfied. Assumption A3 may be interpreted as “blindness” of the measurement errors irrespective of the subject.

Our interest is $\rho_h = \text{cor}(x_i, y_i) = c_{xy} / \sqrt{c_{xx}c_{yy}}$, where $c_{xy} = \text{cov}(x_i, y_i)$, $c_{xx} = \text{var}(x_i)$, and $c_{yy} = \text{var}(y_i)$. Note that the correlation ρ_h , the covariance c_{xy} , and the variances c_{xx} and c_{yy} do not depend on i , because the pairs (x_i, y_i) are i.i.d. (assumption A1); in other words, these are population parameters. Then, we have

$$\text{cov}(x_{is}, y_{is}) = \text{cov}(x_i, y_i) + \text{cov}(\xi_{is}, \eta_{is}) = c_{xy} + c_{\xi\eta}, \quad (3)$$

by A1—A3. Equation (3) shows that the covariance between the repeated measures is not the same as the one between the two variables of interest, but includes another source of covariance (the covariance between the measurement errors). By similar arguments, one has $\text{var}(x_{is}) = c_{xx} + c_{\xi\xi}$ and $\text{var}(y_{is}) = c_{yy} + c_{\eta\eta}$. Thus, we have

$$\text{cor}(x_{is}, y_{is}) = \frac{c_{xy} + c_{\xi\eta}}{\sqrt{c_{xx} + c_{\xi\xi}} \sqrt{c_{yy} + c_{\eta\eta}}}. \quad (4)$$

The right side of (4), denoted by ρ , is what one is actually estimating, if the Pearson correlation, r , is blindly used. Note that ρ is most likely not equal to ρ_h , the true correlation that we are interested in. For example, suppose that all the variances are equal to 1, $c_{xy} = 0.9$, and $c_{\xi\eta} = 0.1$. Then, we have $\rho_h = 0.9$, while $\rho = (0.9 + 0.1)/2 = 0.5$. Another special case is when there is no correlation between ξ and η , i.e., $c_{\xi\eta} = 0$. In this case, it is easy to see that $|\rho|$ is strictly less than $|\rho_h|$. Thus, in the latter case, the (blind) Pearson correlation is expected to underestimate in absolute value the true correlation.

On the other hand, by (2) we have

$$\bar{x}_{i\cdot} = x_i + \bar{\xi}_{i\cdot}, \bar{y}_{i\cdot} = y_i + \bar{\eta}_{i\cdot},$$

where $\bar{x}_{i\cdot} = n_i^{-1} \sum_{s=1}^{n_i} x_{is}$, etc., which imply $\text{cov}(x_{i\cdot}, \bar{y}_{i\cdot}) = c_{xy} + \text{cov}(\bar{\xi}_{i\cdot}, \bar{\eta}_{i\cdot})$, by A1 and A3.

Note that $\text{cov}(\bar{\xi}_{i\cdot}, \bar{\eta}_{i\cdot}) = n_i^{-2} \sum_{s,t=1}^{n_i} \text{cov}(\xi_{is}, \eta_{it}) = n_i^{-2} \sum_{s=1}^{n_i} \text{cov}(\xi_{is}, \eta_{is}) = n_i^{-1} c_{\xi\eta}$, by A2.

Thus, we have $\text{cov}(x_{i\cdot}, \bar{y}_{i\cdot}) = c_{xy} + c_{\xi\eta}/n_i$, or

$$c_{\xi\eta} = n_i \{ \text{cov}(\bar{x}_{i\cdot}, \bar{y}_{i\cdot}) - c_{xy} \}. \quad (5)$$

By bringing (5) to (3) we obtain

$$c_{xy} = \frac{n_i}{n_i - 1} \text{cov}(\bar{x}_{i\cdot}, \bar{y}_{i\cdot}) - \frac{\text{cov}(x_{is}, y_{is})}{n_i - 1}. \quad (6)$$

Because (6) holds for any $s = 1, \dots, n_i$, by taking the average over s , we have

$$c_{xy} = \frac{n_i}{n_i - 1} \text{cov}(\bar{x}_{i\cdot}, \bar{y}_{i\cdot}) - \frac{1}{n_i(n_i - 1)} \sum_{s=1}^{n_i} \text{cov}(x_{is}, y_{is}) = \frac{n_i}{n_i - 1} E(\bar{x}_{i\cdot} \bar{y}_{i\cdot}) - \frac{n_i}{n_i - 1} E(\bar{x}_{1\cdot}) E(\bar{y}_{1\cdot}) - \frac{1}{n_i - 1} \text{cov}(x_{11}, y_{11}), \quad (7)$$

because A1—A3 imply $E(x_{i\cdot}) = E(x_{1\cdot})$, $E(\bar{y}_{i\cdot}) = E(\bar{y}_{1\cdot})$ and $\text{cov}(x_{is}, y_{is}) = \text{cov}(x_{11}, y_{11})$.

Finally, because (7) holds for every i , by taking the average over i we get

$$c_{xy} = \frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} E(\bar{x}_{i\cdot} \bar{y}_{i\cdot}) - \left(\frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} \right) E(\bar{x}_{1\cdot}) E(\bar{y}_{1\cdot}) - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) \text{cov}(x_{11}, y_{11}). \quad (8)$$

The first term on the right side of (8) can be expressed as

$$E \left(\frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_{i\cdot} \bar{y}_{i\cdot} \right),$$

which has the following consistent estimator (by removing the expectation sign):

$$s_1 = \frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_{i\cdot} \bar{y}_{i\cdot}.$$

On the other hand, consistent estimators of $E(x_{1\cdot})$, $E(\bar{y}_{1\cdot})$ and $\text{cov}(x_{11}, y_{11})$ are

$$s_2 = m^{-1} \sum_{i=1}^m \bar{x}_{i\cdot}, s_3 = m^{-1} \sum_{i=1}^m \bar{y}_{i\cdot} \text{ and}$$

$$s_4 = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{s=1}^{n_i} x_{is} y_{is} - s_2 s_3,$$

respectively. Thus, a consistent estimator of c_{xy} is obtained by

$$\hat{c}_{xy} = s_1 - \left(\frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} \right) s_2 s_3 - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) s_4 = s_1 - s_2 s_3 - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{s=1}^{n_i} x_{is} y_{is} \right). \quad (9)$$

Similar to (9), $\hat{c}_{xx}$ and $\hat{c}_{yy}$, the estimators of c_{xx} and c_{yy} , respectively, are obtained (with all the letter y replaced by x for c_{xx} , and the other way around for c_{yy} ; see Appendix). Therefore, a consistent estimator of ρ_h is obtained as

$$r_h = \hat{c}_{xy} / \sqrt{\hat{c}_{xx} \hat{c}_{yy}}. \quad (10)$$

3 p-value in testing hypothesis

In some cases, the interest is whether the correlation between the characteristics of interest is “significant”. This is related to testing the hypothesis $H_0 : \rho_h = 0$ vs $H_a : \rho_h \neq 0$ (or $\rho_h > 0$; $\rho_h < 0$). First note that $\rho_h = \text{cor}(x_i, y_i) = 0$ is equivalent to $c_{xy} = \text{cov}(x_i, y_i) = 0$. Thus, the null hypothesis is equivalent to

$$H_0 : c_{xy} = 0. \quad (11)$$

We consider the following t-statistic

$$t_c = \frac{\hat{c}_{xy}}{\text{s.e.}(\hat{c}_{xy})}. \quad (12)$$

Under regularity conditions, (12) has an asymptotic standard normal distribution. Unlike the Pearson correlation, whose test statistic, $r / \sqrt{(n-2)^{-1}(1-r^2)}$, has a student- t distribution with $(n-2)$ degrees of freedom (d.f.) under the null hypothesis that $\rho = 0$, if the data are bivariate normal (e.g., [15]), the normal or z -distribution is used to determine the critical value for (12). This is more attractive because in the presence of repeated measures and possibly non-normal data, it is difficult to determine the effective d.f. as for testing the Pearson correlation. Note that the derivation of the (exact) t -distribution is based on the assumption of independent normal observations. The current case is much more complicated with correlation among the repeated measures (even under the null hypothesis). Thus, the (exact) t -distribution result no longer holds, even if the data is normal. Furthermore, we consider situations where the data may not be normal, so an exact t -distribution under the null hypothesis cannot be expected. As noted earlier, our case may be viewed as a special case of the linear mixed models. In general, it is difficult to derive an exact t , χ^2 or F distribution for a testing problem in linear mixed models with unbalanced data (and hence determine the corresponding d.f.), even if the data is normal (e.g., [13], sec. 2.1). See section 6 for further discussion on the p-value.

Thus, the p-value for testing (11) can be obtained, provided that we can determine the denominator of (12). Define $u_1 = s_2$, $u_2 = s_3$, $u_3 = s_1$ and

$$u_4 = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{t=1}^{n_i} x_{is} y_{is}.$$

Then, we have $\hat{c}_{xy} = u_3 - u_1 u_2 - c_1 u_4 = g(u_1, u_2, u_3, u_4)$, where

$$c_1 = \frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1}.$$

Using the delta-method (e.g., [16]), we obtain

$$\text{var}(\hat{c}_{xy}) \approx \frac{\partial g}{\partial u'} \text{Var}(u) \frac{\partial g}{\partial u}, \quad (13)$$

where $u = (u_1, u_2, u_3, u_4)'$. Furthermore, we have

$$\frac{\partial g}{\partial u} = (-u_2, -u_1, 1, -c_1)' \quad (14)$$

(and g'/u' is the transpose of g'/u). An estimate of $\text{Var}(u)$, $\hat{V}$, is derived in Appendix. By combining these results, we obtain

$$\text{s.e.}(\hat{c}_{xy}) = \sqrt{\text{the right side of (13) with } \text{Var}(u) \text{ replaced by } \hat{V}}. \quad (15)$$

More generally, one may be interested in testing whether the correlation is at or below a given threshold, that is, $H_0 : \rho_h = \rho_{h0}$ vs $H_a : \rho_h \neq \rho_{h0}$ ($\rho_h > \rho_{h0}$; $\rho_h < \rho_{h0}$), where ρ_{h0} is a given number. The above approach, unfortunately, does not apply to this case (note that the equivalence of $\rho_h = 0$ to $c_{xy} = 0$ simplifies the problem, but there is no such simplification when $\rho_h \neq 0$). A large-sample t -test for the general hypothesis is based on

$$t_r = \frac{r_h - \rho_{h0}}{\text{s.e.}(r_h)}, \quad (16)$$

where $\text{s.e.}(r_h) = \hat{v}$ and the expression of $\hat{v}^2$ is given in Appendix. To get some idea about the accuracy of this s.e. estimator, or equivalently that of $\hat{v}^2$ as the variance estimator, some small simulations are run under the scenario described in the first paragraph of the next section. Note that $\hat{v}^2$ is supposed to satisfy $d = m\{E(\hat{v}^2) - \text{var}(r_h)\} \rightarrow 0$ as $m \rightarrow \infty$. Based on $K = 1000$ simulation runs, we have $d = -0.202$ for $m = 50$; and $d = 0.057$ for $m = 100$. The asymptotic distribution of t_r under the null hypothesis is, again, the standard normal distribution. The approximate power of the test can be obtained by the standard derivation. For example, for testing $H_0 : \rho_h = \rho_{h0}$ against $H_a : \rho_h > \rho_{h0}$, let ρ_h be the true hidden correlation which is greater than ρ_{h0} . Then, at the α -level of significance, the power of test (16) at the alternative ρ_h is approximately $1 - \Phi(z_\alpha - (\rho_h - \rho_{h0})/\text{s.e.}(r_h))$, where $\Phi(\cdot)$ is the cumulative standard normal distribution function and z_α is the corresponding α critical value. The associated $100(1 - \alpha)\%$ confidence interval for ρ_h is

$$[r_h - z_{\alpha/2} \text{s.e.}(r_h), r_h + z_{\alpha/2} \text{s.e.}(r_h)]. \quad (17)$$

More simulation results are presented in the next section.

4 Simulation studies

The simulations are carried out under the following setting: $E(x_i) = E(y_i) = 0$, $\text{var}(x_i) = \text{var}(y_i) = 1$, $\text{cor}(x_i, y_i) = \rho_h$, $E(\xi_{is}) = E(\eta_{is}) = 0$, $\text{var}(\xi_{is}) = \text{var}(\eta_{is}) = 1$, and $\text{cor}(\xi_{is}, \eta_{is}) = \delta$. Unless specifically indicated, the data is generated under the normal distribution. First consider estimation of ρ_h with $\delta = 0.1$. We consider three different true values of ρ_h : 0.3, 0.6 and 0.9. In each case, three different sample sizes are considered: $m = 50$, $m = 100$ and $m = 200$. In all cases, $n_i = 5$, $1 \leq i \leq m$. The Monte Carlo (MC) mean and standard error (s.e.) of the proposed estimator, r_h , based on $K = 1000$ simulation runs, are reported in Table 1. Note that each mean reported in Table 1 is the average of K MC simulation results (as an approximation to the corresponding true expectation of the estimator). Thus, the MC s.e. (of the MC mean) is the sample standard deviation of the K results (MC s.d.) divided by $\sqrt{K}$. So, in particular, from the MC s.e. one can easily obtain the MC s.d. of the estimator. Overall, the estimator appears to perform very well and show sign of consistency as m increases.

The next simulation shows how blindly using the Pearson correlation can be misleading. Because all the variances are equal to one in our simulations, according to (4), the Pearson correlation coefficient is expected to converge (e.g., in probability) to $(c_{xy} + c_{\xi\eta})/2 = \rho$. The second row of Table 2 shows the values of ρ when the true value of $\rho_h (= c_{xy}$ in this case) is 0.6 and the true value of $\delta (= c_{\xi\eta}$ in this case) is $-0.4, -0.2, 0, 0.2, 0.4, 0.6$, respectively. It is seen that ρ is not equal to ρ_h even if $\delta = 0$, that is, when the measurement errors ξ_{is} and η_{is} are uncorrelated. In fact, the only case that $\rho = \rho_h$ is when the correlation between ξ and η is the same as that between x and y , that is, 0.6. We next carry out a simulation study with $m = 100$ and, again, $n_i = 5$, $1 \leq i \leq m$. Here the rows $E(r_h)$ and $E(r)$ correspond to the MC means (MC s.e.'s in the next row) of r_h and r , respectively, where r is the (blind) Pearson correlation coefficient. Again, the results reported in (the last four rows of) Table 2 are based on $K = 1000$ simulation runs. It is seen that the proposed estimator, r_h , of the true hidden correlation, is doing (much) better for all except the last case, where $\rho = 0.6$, compared to the blind Pearson correlation coefficient, r (note that the smaller MC s.e. of r does not really matter when the estimator is seriously biased). In the last case, of course, r is doing better, but this depends on some luck (that δ happens to be 0.6, so that ρ is equal to ρ_h) and one cannot count on this in practice.

As noted (see section 1), the hidden correlation may be estimated by assuming a linear mixed effects model. Write $\alpha_i = (x_i - \mu_x, y_i - \mu_y)'$, where $\mu_x = E(x_i)$ and $\mu_y = E(y_i)$. Then, (2) can be expressed as $Y_{is} = \mu + \alpha_i + \varepsilon_{is}$, $i = 1, \dots, m$, $s = 1, \dots, n_i$, where $Y_{is} = (x_{is}, y_{is})'$, $\mu = (\mu_x, \mu_y)'$, and $\varepsilon_{is} = (\xi_{is}, \eta_{is})'$. The latter may be viewed as a linear mixed model (e.g., [13]) with the α_i 's being the random effects. Furthermore, assume that $\alpha_1, \dots, \alpha_m$ are independent and bivariate normal with means zero, variances σ_1^2, σ_2^2 , and correlation coefficient ρ_h , and the ε_{is} 's are independent and bivariate normal with means zero, variances τ_1^2, τ_2^2 , and correlation coefficient δ . Then, $\sigma_1^2, \sigma_2^2, \rho_h, \tau_1^2, \tau_2^2$ and δ are unknown variance components. In the current case, because $n_i = 5$, $1 \leq i \leq m$, the model may be further expressed as $Y_i = X_i\mu + Z_i\alpha_i + \varepsilon_i$, $i = 1, \dots, m$, where $Y_i = (Y_{is})_{s=1}^5$, $X_i = Z_i = 1_5 \otimes I_2$ [here $\otimes$ denotes the Kronecker product (e.g., [13], Appendix B.1), $1_5 = (1, 1, 1, 1, 1)'$ and I_2 is the 2×2 identity matrix], and $\varepsilon_i = (\varepsilon_{is})_{s=1}^5$; or, in the standard matrix expression, $Y = X\mu + Z\alpha + \varepsilon$, where $Y = (Y_i)_{i=1}^m$, $X = (X_i)_{i=1}^m$, $Z = \text{diag}(Z_i, 1 \leq i \leq m)$, $\alpha = (\alpha_i)_{i=1}^m$, and $\varepsilon = (\varepsilon_i)_{i=1}^m$. Even in this simple case, the variance-covariance structure is not simple with 6 unknown variance components involved (of which only ρ_h is of interest). The implementation of the procedure in SAS PROC MIXED is not straightforward, and requires understanding of the above linear

mixed model, somehow in a more complicated way as described in [14]. Nevertheless, we are able to compare our method with that of [14]. Two cases are considered. In the first case, the characteristics of interest and measurement errors are both normally distributed. This is the ideal situation for REML. In the second case, the characteristics of interest both have a normal-mixture distribution that consists of 90% of $N(4, 0.1)$ and 10% of $N(-4, 0.1)$. This is the same distribution considered by [17], with mean 3.2 and variance 5.86. The distribution is then standardized to satisfy the requirements set up at the beginning of the section. In addition, the measurement errors have a centralized exponential distribution, that is, the distribution of $X - 1$, where X has the Exponential distribution with mean equal to 1. Two different sample sizes are considered: $m = 50$ and $m = 100$. Results of same-data comparison based on $K = 100$ simulations are reported in Tables 3 and 4. The results are remarkably close. In fact, there is hardly any consistent pattern in favor of any method in terms of the estimation performance. On the other hand, there is a significant difference in terms of the computational efficiency. For example, for the normal case with $m = 50$, $\rho_h = 0.3$, the SAS PROC MIXED procedure took roughly 35 times of the time it took for our procedure to run the same simulations when the jobs were performed on a Dell OPTIPLEX 755 desktop computer (3.0 Ghz Intel E8400 CPU; 3.25 Gb RAM).

Next, we present some simulation results regarding testing hypotheses for the hidden correlation. First we consider $H_0 : \rho_h = 0$ vs $H_a : \rho_h > 0$. We consider both a relatively large sample size, $m = 100$, and a relatively small one, $m = 20$. The results based on $K = 1000$ simulation runs are reported in Table 5, where α is the level of significance, the column $\rho_h = 0$ corresponds to the size of the test, and the columns $\rho_h = 0.3$ and $\rho_h = 0.6$ to the powers at the given alternatives. The value of δ is equal to 0.1, as in Table 1. It appears that the t -test performs quite well in terms of the size, even with the relatively sample size ($m = 20$). On the other hand, the powers are much lower for $m = 20$ than for $m = 100$, as expected. Next, we consider testing $H_0 : \rho_h = 0.5$ vs $H_a : \rho_h > 0.5$. We consider, in addition to $m = 20$ and $m = 100$, a larger sample size $m = 200$. The results based on $K = 1000$ simulation runs are reported in Table 6. It appears that, with the smaller sample size, the test somewhat over-rejects under the null hypothesis, especially for $m = 20$, but the accuracy picks up quickly as the sample size increases. The power follows a reasonable trend as in the previous case.

Finally, we present some simulation results that compare our method with REML in terms of the testing performance. We consider testing the hypothesis $H_0 : \rho_h = 0$ vs $H_a : \rho_h \neq 0$. We compare our test (12) with the REML-based likelihood-ratio test (LRT). The LRT is well-known in the statistical literature. For testing a simple null hypothesis vs a simple alternative, the fundamental Neyman-Pearson lemma (e.g., [18]) states that the LRT is most powerful among all tests of the same significance level. For a discussion of the LRT in linear mixed models, see [13] (sec. 2.1). We consider the case of normal data. The setup is considered most favorable for the LRT. Three different sample sizes are considered: $m = 20$, 50 and 100, ranging from small to relatively large. We consider the same null and alternatives as in Table 5. The numbers of simulation runs for $m = 20, 50, 100$ are $K = 500, 200, 100$, respectively (more simulation runs are needed for smaller sample size in order to obtain stable results). The results are reported in Table 7. It is seen that, overall, the LRT performs somewhat better, but the difference is diminishing as the sample size increases. On the other hand, our method is, again, much more favorable in terms of the computational efficiency.

5 Home care worker example revisited

Now let us return to the home care worker study discussed in section 1. We apply our method to estimate the hidden correlation between each pair of variables. Partial results of the estimated correlation coefficients, r_h , are reported in Table 8 with the corresponding p-

values in the parentheses. As noted, the Pearson correlation coefficient is inappropriate for these hidden correlation problems. Nevertheless, it might be interested in knowing what one is getting if the Pearson method is blindly used. The results show that the Pearson correlation coefficient, r , is, in general, lower in absolute value than our estimate, r_h , which is consistent with our expectation [see the end of the discussion following (4)]. Table 9 provides a summary of these comparisons, where reported are the numbers of estimated correlation coefficients whose absolute values are greater than or equal to the cut-off, for both methods. It is clear that the r_h estimates are, in general, higher than the r estimates in absolute values. Recall that r may be serious biased (see Table 2). As examples of specific comparisons, the r corresponding to the first row of Table 8 are: 1.00 (NA), 0.33 (0.00), 0.20 (0.00), 0.38 (0.00), 0.47 (0.00), with the p-values in the parentheses. It should be noted that the p-value for the Pearson correlation may be misleading in this case, because it is computed based on the assumption that all the repeated measures are independent. However, because the observations are clustered, there may be within-cluster correlations. The implication is that the effective sample size (used to determine the standard errors) is not the same as the total sample size, but is usually smaller. Therefore, in the bottom row of Table 9 we only report the numbers of significant r_h results (at 5% level) that are greater than or equal to the cut-off in absolute values.

Significant highly correlated ($r_h \geq 0.8$) variables include Heavy Object Time & Pain Lag, Heavy Object Time & Non-lagged Pain, Driving for Client & House Clean Time, Bathing/Toileting Time & Dressing Time, House Clean Time & Task Minute Sum, Heavy Object Time & Task Minute Sum, Daily Occupational Fatigue & Daily Negative Affect, and so on. These can be easily interpreted, and agree quite well with the common sense.

6 Discussion

We have developed a simple, Pearson-type, estimator of the correlation coefficient between unobservable characteristics of interest, the hidden correlation. Our method is developed without making serious assumptions about the variances/covariance of the measurement errors, such that $c_{\xi\eta} = 0$, or the normality assumption about the data. This makes the method more applicable to practical situations. Furthermore, we have derived large sample standard errors associated with the hidden correlation estimator, which can be used for hypothesis testing and confidence intervals. Simulation results suggest that the proposed estimator works well in both estimation and testing performances, especially when the sample size is relatively large. Furthermore, we show by a simulation study that the proposed simple estimator performs similarly to the estimator obtained via the more sophisticated mixed-effects model approach, but our method is much more efficient in terms of computational speed. A potential limitation of our method is that its testing procedure relies on large sample approximations, which we discuss in further details below.

The p-value calculation (see section 3) is based on the asymptotic normal distribution. However, the normal approximation may not be accurate when the sample size is small. For example, it is seen in Table 6 that the simulated size is much higher than the nominal level when the sample size is small ($m = 20$). In such a case, permutation or bootstrap methods are often used in other contexts of studies. Permutation tests rely on exchangeability under the null hypothesis. In the current case, however, the null hypothesis does not seem to lead to useful exchangeability. For example, it is not clear at all if any exchangeability can be implied under the null hypothesis $H_0 : \rho_h = 0.3$. In fact, even $H_0 : \rho_h = 0$ does not seem helpful, unless it is known in advance that $c_{\xi\eta} = 0$ (see below for detail).

It also seems difficult to bootstrap the p-value in the current situation. To see this, first note that this is not a conventional bootstrap based on the observed data—the null hypothesis has

to be involved in the resampling. The simplest case would be to test $H_0 : \rho_h = 0$. If the x_i 's and y_i 's are directly observable, then a non-parametric bootstrap can be done by drawing samples $x_1^*, \dots, x_m^*$ and $y_1^*, \dots, y_m^*$ independently, and then pairing x_i^* with y_i^* , $i = 1, \dots, m$. Note that the null hypothesis does play a role in this resampling procedure [otherwise, it is not equivalent to resampling (x_i^*, y_i^*) , $i = 1, \dots, m$ from $\{(x_i, y_i), i = 1, \dots, m\}$]. However, this strategy does not work if the x_i 's and y_i 's are not directly observable. Furthermore, a similar approach does not work with the x_i, y_i replaced by x_{is}, y_{is} . This is because, even if x_i, y_i are independent (under the null hypothesis), there is no independence between x_{is} and y_{is} ; in fact, $\text{cov}(x_{is}, y_{is}) = c_{\xi\eta}$ under the null hypothesis, and the latter may not be zero.

Another difficulty of the non-parametric bootstrap is due to the fact that the n_i 's are not necessarily equal. Note that, if the n_i 's are the same, say, $n_i = k$, $1 \leq i \leq m$, then the vectors $(x_{is}, y_{is}, s = 1, \dots, k)$, $i = 1, \dots, m$ are i.i.d. However, assuming that the n_i 's are equal may not be practical. Even in the special case with all the n_i 's equal, it is still unclear how the null hypothesis can come into play in a practical way.

On the other hand, it is often easier to incorporate the null hypothesis with the model-based, or parametric bootstrap. Once again, consider the simple case of testing $H_0 : \rho_h = 0$. If normality is assumed, one could estimate all the unknown parameters except ρ_h , which is set to 0 under H_0 , and then use the estimated parameters and $\rho_h = 0$ (as the true parameters) to generate the data. However, this approach is problematic in our case. First, because a parametric model is not assumed, the parametric-bootstrap approach does not work. Second, even if one is willing to make stronger assumptions, such as normality, the accuracy of the parameter estimation is, again, dependent on the sample size. Recall that the bootstrap idea is brought about due to the concern of small sample size that results in inaccurate normal approximation. The very same concern arises with the parameter estimation, which results in inaccurate bootstrapping. As written by Efron [19], bootstrap provides "approximate frequency statements, not approximate likelihood statements. Fundamental inference problems remain, no matter how well the bootstrap works."

In a way, the nature of difficulty in bootstrapping encountered here is similar to that under a mixed-effects model. Note that the latter also involve unobservable random quantities (the random effects) that cannot be directly bootstrapped. See, for example, [20] and [21] for some recent developments. The strategies in the latter studies will be explored in our future study. It would also be interesting to explore some other more advanced bootstrap ideas, such as the *block bootstrap* ([22]).

Acknowledgments

The authors' research is partially supported by the NIH grant R01-GM085205A1. In addition, Jiming Jiang's research is partially supported by the NSF grant DMS-0809127. The authors are grateful to Drs. Ryan Olson and Brad Wipfli of the Center for Research on Occupational and Environmental Toxicology at the Oregon Health & Science University for providing the data from their research as well as information and helpful discussions. The authors are grateful to an Associate Editor and two referees for their thoughtful comments that led to improvement of the manuscript.

Appendix

A.1 Complete analytic expressions of $\hat{\mathbf{c}}_{xy}$, $\hat{\mathbf{c}}_{xx}$, $\hat{\mathbf{c}}_{yy}$

$$\begin{aligned}\hat{c}_{xy} &= \frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_{i.} \bar{y}_{i.} - \left(\frac{1}{m} \sum_{i=1}^m \bar{x}_{i.} \right) \left(\frac{1}{m} \sum_{i=1}^m \bar{y}_{i.} \right) - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{s=1}^{n_i} x_{is} y_{is} \right); \\ \hat{c}_{xx} &= \frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} (\bar{x}_{i.})^2 - \left(\frac{1}{m} \sum_{i=1}^m \bar{x}_{i.} \right)^2 - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{s=1}^{n_i} x_{is}^2 \right); \\ \hat{c}_{yy} &= \frac{1}{m} \sum_{i=1}^m \frac{n_i}{n_i - 1} (\bar{y}_{i.})^2 - \left(\frac{1}{m} \sum_{i=1}^m \bar{y}_{i.} \right)^2 - \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i - 1} \right) \left(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i} \sum_{s=1}^{n_i} y_{is}^2 \right).\end{aligned}$$

A.2 Derivation of $\hat{\mathbf{V}}$

Define

$$c_2 = \frac{1}{m} \sum_{i=1}^m \left(\frac{n_i}{n_i - 1} \right)^2, \quad \hat{v}_{11} = \sum_{i=1}^m \bar{x}_{i.}^2 - m * u_1^2, \quad \hat{v}_{22} = \sum_{i=1}^m \bar{y}_{i.}^2 - m * u_2^2$$

$[\bar{x}_{i.}^2 = (x_{i.})^2$, etc.]. Now, let

$$c_2 = \frac{1}{m} \sum_{i=1}^m \left(\frac{n_i}{n_i - 1} \right)^2, \quad u_5 = \frac{1}{m} \sum_{i=1}^m \bar{x}_{i.} \bar{y}_{i.}.$$

We continue with the definitions of $\hat{v}$'s:

$$\begin{aligned}\hat{v}_{33} &= \sum_{i=1}^m \left(\frac{n_i}{n_i - 1} \right)^2 \bar{x}_{i.}^2 \bar{y}_{i.}^2 - m * c_2 * u_5^2, \\ \hat{v}_{44} &= \sum_{i=1}^m \left(\frac{1}{n_i} \sum_{s=1}^{n_i} x_{is} y_{is} \right)^2 - m * u_4^2, \\ \hat{v}_{12} &= \sum_{i=1}^m \bar{x}_{i.} \bar{y}_{i.} - m * u_1 * u_2, \\ \hat{v}_{13} &= \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_{i.}^2 \bar{y}_{i.} - m(c_1 + 1) * u_1 * u_5.\end{aligned}$$

Next, define $z_{is} = x_{is} y_{is}$, $\bar{z}_{i.} = n_i^{-1} \sum_{s=1}^{n_i} z_{is}$, and $u_6 = m^{-1} \sum_{i=1}^m \bar{z}_{i.}$. We define

$$\begin{aligned}\hat{v}_{14} &= \sum_{i=1}^m \bar{x}_i \bar{z}_i - m * u_1 * u_6, \\ \hat{v}_{23} &= \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_i \bar{y}_i^2 - m(c_1 + 1) * u_2 * u_5, \\ \hat{v}_{24} &= \sum_{i=1}^m \bar{y}_i \bar{z}_i - m * u_2 * u_6, \\ \hat{v}_{34} &= \sum_{i=1}^m \frac{n_i}{n_i - 1} \bar{x}_i \bar{y}_i \bar{z}_i - m(c_1 + 1) * u_5 * u_6.\end{aligned}$$

Extend the definition of $\hat{v}$ according to symmetry, that is, $\hat{v}_{kl} = \hat{v}_{lk}$, $k, l = 1, \dots, 4$. Then, $\text{Var}(u)$ in (13) is estimated by $m^{-2}V = m^{-2}(\hat{v}_{kl})_{k,l=1, \dots, 4}$.

A.3 Derivation of s.e.(r_h)

The derivation here is “semirigorous”. For example, we shall assume $E\{o_P(m^{-1})\} = o(m^{-1})$, although some regularity conditions are needed for this to hold. We refer the details to a technical report (see [23]). First, it is easy to show, by the Taylor expansion and the facts that $\hat{c}_{xy} - c_{xy} = O_P(m^{-1/2})$, etc., we have

$$r_h - \rho_h = \frac{\hat{c}_{xy} - c_{xy}}{\sqrt{c_{xx}c_{yy}}} - \frac{c_{xy}}{2\sqrt{c_{xx}c_{yy}}} \left(\frac{\hat{c}_{xx} - c_{xx}}{c_{xx}} + \frac{\hat{c}_{yy} - c_{yy}}{c_{yy}} \right) + o_P(m^{-1/2}).$$

Furthermore, it can be shown that

$$\begin{aligned}\hat{c}_{xy} - c_{xy} &= m^{-1} \sum_{i=1}^m \xi_{i,xy} + o_P(m^{-1/2}), \quad \hat{c}_{xx} - c_{xx} = m^{-1} \sum_{i=1}^m \xi_{i,xx} + o_P(m^{-1/2}), \text{ and} \\ \hat{c}_{yy} - c_{yy} &= m^{-1} \sum_{i=1}^m \xi_{i,yy} + o_P(m^{-1/2}), \text{ where}\end{aligned}$$

$$\begin{aligned}\xi_{i,xy} &= \frac{n_i}{n_i - 1} \bar{x}_i \bar{y}_i - E(y_{11}) \bar{x}_i - E(x_{11}) \bar{y}_i - E(x_{11})E(y_{11}) - \frac{c_1}{n_i} \sum_{s=1}^{n_i} x_{is} y_{is} - c_{xy} - \left(\frac{1}{n_i - 1} - c_1 \right) E(x_{11} y_{11}), \\ \xi_{i,xx} &= \frac{n_i}{n_i - 1} \bar{x}_i^2 - 2E(x_{11}) \bar{x}_i - \{E(x_{11})\}^2 - \frac{c_1}{n_i} \sum_{s=1}^{n_i} x_{is}^2 - c_{xx} - \left(\frac{1}{n_i - 1} - c_1 \right) E(x_{11}^2),\end{aligned}$$

and $\xi_{i,yy}$ is $\xi_{i,xx}$ with x replaced by y . Note that all the ξ 's have mean zero. Thus, combining the results, we have $r_h - \rho_h = m^{-1} \sum_{i=1}^m \xi_i + o_P(m^{-1/2})$, where

$$\xi_i = \frac{1}{\sqrt{c_{xx}c_{yy}}} \left\{ \xi_{i,xy} - \frac{c_{xy}}{2} \left(\frac{\xi_{i,xx}}{c_{xx}} + \frac{\xi_{i,yy}}{c_{yy}} \right) \right\}.$$

Note that ξ_i , $i = 1, \dots, m$ are independent with mean 0. Thus, we have

$$\begin{aligned}\text{var}(r_h) &= m^{-2} E\left(\sum_{i=1}^m \xi_i^2\right) + o(m^{-1}). \text{ A first-order approximation to } \text{var}(r_h), \text{ in the sense that} \\ &\text{the difference is } o(m^{-1}) \text{ [note that } \text{var}(r_h) = O(m^{-1}) \text{] (e.g., [24], sec. 4.8), is given by} \\ \hat{v}^2 &= m^{-2} \sum_{i=1}^m \hat{\xi}_i^2, \text{ where } \hat{\xi}_i \text{ is } \xi_i \text{ with } c_{xy}, c_{xx}, c_{yy}, E(x_{11}), E(y_{11}), E(x_{11}y_{11}), E(x_{11}^2), E(y_{11}^2) \\ &\text{replaced by } \hat{c}_{xy}, \hat{c}_{xx}, \hat{c}_{yy},\end{aligned}$$

$m^{-1}\sum_{i=1}^m \bar{x}_i, m^{-1}\sum_{i=1}^m \bar{y}_i, m^{-1}\sum_{i=1}^m n_i^{-1}\sum_{s=1}^{n_i} x_{is}y_{is}, m^{-1}\sum_{i=1}^m n_i^{-1}\sum_{s=1}^{n_i} x_{is}^2, m^{-1}\sum_{i=1}^m n_i^{-1}\sum_{s=1}^{n_i} y_{is}^2$, respectively. Therefore, we have the expression $s.e.(r_h) = \hat{v} = m^{-1} \sqrt{\sum_{i=1}^m \hat{\xi}_i^2}$.

References

1. Rodgers JL, Nicewander WA. Thirteen ways to look at the correlation coefficient. *The American Statistician*. 1988; 42:59–66.
2. Bland JM, Altman DG. Statistics notes: measurement error and correlation coefficient. *BMJ*. 1996; 313:41. [PubMed: 8664775]
3. Fisher, RA. *Statistical Methods for Research Workers*. 12th ed. Oliver & Boyd; 1954.
4. Koch, GG. Intraclass correlation coefficient. In: Kotz, S.; Johnson, NL., editors. *Encyclopedia of Statistical Sciences*. Vol. 4. New York: Wiley; 1982. p. 213-217.
5. Graybill, FA. *An Introduction to Linear Statistical Models*, Vol I. New York: McGraw-Hill; 1961.
6. Horn M, Guiard V. Correlation and linear regression if the random variables are subject to errors or fluctuations. *Biom. J*. 1986; 28:683–696.
7. Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: Implications for study design and hypothesis testing. *Amer. J. Epidemiol*. 1988; 127:377–386. [PubMed: 3337089]
8. Rifkin RD. Effects of correlated and uncorrelated measurement error on linear regression and correlation in medical method comparison studies. *Statist. Med*. 1995; 14:789–798.
9. Archer KJ, Dumur CI, Taylor GS, Chaplin MD, Guiseppi-Elie A, Buck GA, Grant G, Ferreira-Gonzalez A, Gurrett CT. A disattenuated correlation estimate when variables are measured with error: Illustration estimating cross-platform correlations. *Statist. Med*. 2008; 27:1026–1039.
10. Schaalje GB, Butts RA. Some effects of ignoring correlated measurement errors in straight line regression and prediction. *Biometrics*. 1993; 49:1262–1267.
11. Thoresen M, Laake P. On the simple linear regression model with correlated measurement errors. *J. Statist. Planning Inf*. 2007; 137:68–78.
12. Anuradha R. Estimating correlation coefficient between two variables with repeated observations using mixed effects model. *Biom. J*. 2006; 48:286–301. [PubMed: 16708779]
13. Jiang, J. *Linear and Generalized Linear Mixed Models and Their Applications*. New York: Springer; 2007.
14. Hamlett A, Ryan L, Wolfinger R. *SUGI Proc. Statist. Data Anal*. 2003; 198-29:1–7.
15. Rahman, NA. *A Course in Theoretical Statistics*. Charles Griffin & Comp; 1968.
16. Lehmann, EL. *Elements of Large-Sample Theory*. New York: Springer; 1999.
17. Claeskens G, Hart J. Goodness-of-fit tests in mixed models (with discussion). *TEST*. 2009; 18:213–270.
18. Lehmann, EL. *Testing Statistical Hypothesis*. 2nd ed.. New York: Springer; 1986.
19. Efron B. Bootstrap method: Another look at the jackknife. *Ann. Statist*. 1979; 7:1–26.
20. Hall P, Maiti T. Nonparametric estimation of mean-squared prediction error in nested-error regression models. *Ann. Statist*. 2006; 34:1733–1750.
21. Chatterjee S, Lahiri P, Li H. Parametric bootstrap approximation to the distribution of EBLUP, and related prediction intervals in linear mixed models. *Ann. Statist*. 2008; 36:1221–1245.
22. Künsch HR. The jackknife and the bootstrap for general stationary observations. *Ann. Statist*. 1989; 17:1217–1241.
23. Nguyen, T.; Jiang, J. Tech. Report. Davis, CA: Dept. Statist., Univ. Calif.; 2011. First-order approximation to the variance of the hidden correlation estimator.
24. Jiang, J. *Large Sample Techniques for Statistics*. New York: Springer; 2010.

Table 1

Monte Carlo Mean and S.E. of Estimator

<i>m</i>	$\rho_h = 0.3$		$\rho_h = 0.6$		$\rho_h = 0.9$	
	mean	s.e.	mean	s.e.	mean	s.e.
50	0.301	0.005	0.600	0.004	0.907	0.002
100	0.296	0.004	0.598	0.003	0.903	0.001
200	0.298	0.003	0.600	0.002	0.900	0.001

Table 2

Comparison with Pearson Correlation: True $\rho_h = 0.6$; E—Monte Carlo Mean; s.e.—Monte Carlo S.E.

δ	-0.4	-0.2	0	0.2	0.4	0.6
ρ	0.1	0.2	0.3	0.4	0.5	0.6
$E(r_h)$	0.605	0.603	0.601	0.601	0.597	0.595
s.e.(r_h)	0.003	0.003	0.003	0.003	0.003	0.002
$E(r)$	0.096	0.197	0.298	0.400	0.498	0.599
s.e.(r)	0.002	0.002	0.002	0.002	0.002	0.001

Table 3

Comparison with REML—Normal Case: MC Mean (S.E.), % Bias (S.E.)

<i>m</i>	True ρ_h	r_h				REML			
		mean	% bias	s.e.	% s.e.	mean	% bias	s.e.	% s.e.
50	0.3	0.306	2.1	0.017	5.5	0.306	2.1	0.015	5.1
50	0.6	0.600	0.0	0.013	2.2	0.596	-0.6	0.013	2.2
50	0.9	0.900	0.0	0.007	0.7	0.896	-0.5	0.006	0.7
100	0.3	0.311	3.6	0.009	3.2	0.311	3.7	0.009	3.1
100	0.6	0.618	3.0	0.007	1.2	0.615	2.5	0.008	1.3
100	0.9	0.898	-0.2	0.004	0.5	0.897	-0.3	0.004	0.5

Table 4

Comparison with REML—Non-normal Case: MCMEan (S.E.), % Bias (S.E.)

<i>m</i>	True ρ_h	r_h				REML			
		mean	% bias	s.e.	% s.e.	mean	% bias	s.e.	% s.e.
50	0.3	0.319	6.2	0.019	6.2	0.320	6.7	0.018	6.0
50	0.6	0.599	-0.2	0.017	2.8	0.591	-1.5	0.016	2.7
50	0.9	0.904	0.4	0.009	1.1	0.899	-0.1	0.009	1.0
100	0.3	0.296	-1.4	0.011	3.7	0.297	-1.0	0.011	3.7
100	0.6	0.622	3.6	0.009	1.5	0.621	3.5	0.009	1.5
100	0.9	0.900	0.0	0.005	0.6	0.898	-0.2	0.005	0.6

Table 5

Size and Power of t -test ($H_0 : \rho_h = 0$)

α	$m = 20$			$m = 100$		
	$\rho_h = 0$	$\rho_h = 0.3$	$\rho_h = 0.6$	$\rho_h = 0$	$\rho_h = 0.3$	$\rho_h = 0.6$
0.10	0.118	0.491	0.873	0.093	0.878	1.000
0.05	0.057	0.321	0.761	0.050	0.799	1.000

Table 6

Size and Power of t -test ($H_0 : \rho_h = 0.5$)

m	$\alpha = 0.10$			$\alpha = 0.05$		
	$\rho_h = 0.5$	$\rho_h = 0.7$	$\rho_h = 0.9$	$\rho_h = 0.5$	$\rho_h = 0.7$	$\rho_h = 0.9$
20	0.207	0.589	0.942	0.161	0.497	0.918
100	0.144	0.898	1.000	0.097	0.838	1.000
200	0.118	0.986	1.000	0.058	0.970	1.000

Table 7

Comparison with (REML) Likelihood-Ratio Test									
<i>m</i>	α	$\rho_h = 0$		$\rho_h = 0.3$		$\rho_h = 0.6$		r_h	LRT
		r_h	LRT	r_h	LRT	r_h	LRT		
20	0.10	0.142	0.106	0.334	0.314	0.748	0.798		
20	0.05	0.066	0.046	0.232	0.232	0.592	0.692		
50	0.10	0.105	0.070	0.555	0.585	0.960	0.975		
50	0.05	0.055	0.050	0.405	0.425	0.930	0.955		
100	0.10	0.110	0.140	0.830	0.840	1.000	1.000		
100	0.05	0.080	0.070	0.770	0.770	1.000	1.000		

Table 8

Estimated Correlations and P-values: Partial results of the estimated correlation coefficient, r_h , with the corresponding p-value in the parentheses.

Variable	3	4	5	6	7
3	1.00 (NA)	0.62 (0.11)	0.63 (0.05)	0.51 (0.01)	0.94 (0.25)
4		1.00 (NA)	0.84 (0.00)	0.52 (0.01)	0.69 (0.02)
5			1.00 (NA)	0.45 (0.00)	0.71 (0.01)
6				1.00 (NA)	0.41 (0.00)
7					1.00 (NA)

Table 9

# of Absolute Values of Estimates		Cut-off									
	Cut-off	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	
r_h		220	176	141	97	62	43	28	17	8	
r		169	95	63	37	16	7	4	1	0	
Significant r_h		47	45	43	40	30	21	12	9	4	