Published in final edited form as:

Stat Med. 2017 June 30; 36(14): 2163–2186. doi:10.1002/sim.7257.

Estimation of Rank Correlation for Clustered Data

Bernard Rosner¹ and Robert Glynn²

¹Professor of Medicine/Biostatistics, Harvard Medical School, Channing Division of Network Medicine, 181 Longwood Avenue, Boston, MA 02115

²Professor of Medicine, Harvard Medical School, Department of Medicine, Division of Preventive Medicine, 900 Commonwealth Avenue, Boston, MA 02115, National Eye Institute, R01 EY022445

Abstract

It is well known that the sample correlation coefficient (R_{xy}) is the maximum likelihood estimator (MLE) of the Pearson correlation (ρ_{xy}) for i.i.d. bivariate normal data. However, this is not true for ophthalmologic data where X (e.g., visual acuity) and Y (e.g., visual field) are available for each eye and there is positive intraclass correlation for both X and Y in fellow eyes. In this paper, we provide a regression-based approach for obtaining the MLE of ρ_{xy} for clustered data, which can be implemented using standard mixed effects model software. This method is also extended to allow for estimation of partial correlation by controlling both X and Y for a vector $\underline{U}$ of other covariates. In addition, these methods can be extended to allow for estimation of rank correlation for clustered data by (a) converting ranks of both X and Y to the probit scale, (b) estimating the Pearson correlation between probit scores for X and Y , and (c) using the relationship between Pearson and rank correlation for bivariate normally distributed data. The validity of the methods in finite-sized samples is supported by simulation studies. Finally, two examples from ophthalmology and analgesic abuse are used to illustrate the methods.

Keywords

clustered data; rank correlation; Pearson correlation; partial correlation

1. Introduction

It is well known that the sample correlation coefficient (R_{xy}) is the maximum likelihood estimator (MLE) of the Pearson correlation coefficient (ρ_{xy}) for i.i.d. bivariate normally distributed data.[1] Also, there is an extensive literature on estimation of regression models for clustered normally distributed data.[2] However, estimation of Pearson correlation in the clustered data setting has received less attention. Furthermore, many outcome measures in ophthalmology are either continuous and not normally distributed or are ordinal. For example, the Humphrey visual field is often used to characterize visual field in retinitis pigmentosa (RP) patients. The Humphrey field is continuous, but is not normally distributed (see Figure 1). Similarly, Early Treatment Diabetic Retinopathy Study (ETDRS)[3] visual acuity, a measure of central vision, is also frequently not normally distributed (see Figure 2).

Thus, the rank correlation is a natural measure of association between these two measures. However, both measures are correlated in fellow eyes which violates standard assumptions of inference regarding rank correlation (see Figure 3).

The subject of this paper is to provide methods for obtaining point and interval estimates of rank correlation for clustered data. We first develop methods for estimation of Pearson correlation and then extend these methods to estimation of rank correlation. Methods are first presented for the case of two subunits per cluster and extended to allow for more than two subunits per cluster. Extensions for unbalanced data (i.e., a variable number of subunits per cluster) are also provided.

2. Estimation of Pearson correlation for clustered data

(a) Clusters of size 2

We assume throughout this paper that subunits within a cluster are distinguishable (e.g., right eye (OD) and left eye (OS) scores). We also assume that the Pearson correlation between X and Y scores is the same for right and left eyes. We will begin with a discussion of maximum likelihood estimation of Pearson correlation for clusters of size 2 and will extend our inference to clusters of arbitrary size. Let $X_{ij} = X$ score for the j^{th} subunit in the i^{th} cluster; $i = 1, \dots, n$; $j = 1, 2$ and Y_{ij} defined similarly for the Y score. Our goal is to estimate $\text{corr}(X_{ij}, Y_{ij})$. For simplicity, we will use the notation (x_{ij}, y_{ij}) to denote both the random variable (X_{ij}, Y_{ij}) as well as its sample realization. Suppose we have the data

$$\underline{z}_i = \begin{pmatrix} \underline{x}_i \\ \underline{y}_i \end{pmatrix}, \text{ where } \underline{x}_i = (x_{i1}, x_{i2})', \underline{y}_i = (y_{i1}, y_{i2})'; i = 1, \dots, n.$$

Thus, $\underline{x}_i$ and $\underline{y}_i$ are 2×1 vectors and $\underline{z}_i$ is a 4×1 vector.

We assume that $\underline{z}_i$ is multivariate normal with mean $\underline{\mu} = (\mu_x, \mu_x, \mu_y, \mu_y)'$ and covariance matrix Σ given by

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho_{xx}\sigma_x^2 & \rho_{xy}\sigma_x\sigma_y & \rho_{xy}^*\sigma_x\sigma_y \\ \rho_{xx}\sigma_x^2 & \sigma_x^2 & \rho_{xy}^*\sigma_x\sigma_y & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \rho_{xy}^*\sigma_x\sigma_y & \sigma_y^2 & \rho_{yy}\sigma_y^2 \\ \rho_{xy}^*\sigma_x\sigma_y & \rho_{xy}\sigma_x\sigma_y & \rho_{yy}\sigma_y^2 & \sigma_y^2 \end{pmatrix} \equiv \begin{pmatrix} \sum_x & \sum_{xy} \\ \sum_{yx} & \sum_y \end{pmatrix}$$

Thus, $\rho_{xy} = \text{corr}(x_{ij}, y_{ij})$ = inter-class Pearson correlation between visual acuity (VA) (x_{ij}) and visual field (VF) (y_{ij}) for the same eye, $i = 1, n$; $j = 1, 2$.

$\rho_{xy}^* = \text{corr}(x_{ij_1}, y_{ij_2})$ $\equiv$ cross- correlation between X and Y = inter-class Pearson correlation between VA for the right eye (OD) and VF for the left eye (OS) and conversely, $i = 1, \dots, n$; $j_1, j_2 = 1, 2$.

$\rho_{xx} = \text{corr}(x_{ij_1}, x_{ij_2})$ = intraclass correlation between VA for the left and right eyes, $i = 1, \dots, n$; $j_1, j_2 = 1, 2$.

$\rho_{yy} = \text{corr}(y_{ij1}, y_{ij2})$ = intraclass correlation between VF for the left and right eyes, $i = 1, \dots, n$; $j_1, j_2 = 1, 2$.

The likelihood for the i^{th} subject is given by

$$L_i = \frac{1}{(2\pi)^2 |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} (z_i - \mu)' \Sigma^{-1} (z_i - \mu)\right]$$

and the overall likelihood is given by:

$$L = \prod_{i=1}^n L_i$$

The goal is to obtain maximum likelihood estimates (MLE) of ρ_{xy} . Note that if

$\rho_{xx} = \rho_{yy} = \rho_{xy}^* = 0$, then the MLE of ρ_{xy} is the ordinary Pearson product-moment correlation [1] given by

$$\hat{\rho}_{xy, \text{Pearson}} = \sum_{i=1}^n \sum_{j=1}^2 (x_{ij} - \bar{x})(y_{ij} - \bar{y}) / \left[\sum_{i=1}^n \sum_{j=1}^2 (x_{ij} - \bar{x})^2 \sum_{i=1}^n \sum_{j=1}^2 (y_{ij} - \bar{y})^2 \right]^{1/2}$$

where

$$\bar{x} = \sum_{i=1}^n \sum_{j=1}^2 x_{ij} / (2n)$$

and

$$\bar{y} = \sum_{i=1}^n \sum_{j=1}^2 y_{ij} / 2n$$

However, it is difficult to obtain closed form maximum likelihood estimates of ρ_{xy} using standard software if $\underline{z}_{i1} = (x_{i1}, y_{i1})$ and $\underline{z}_{i2} = (x_{i2}, y_{i2})$ are not independent. Instead, we consider an indirect approach. Specifically, if we factor L_i as follows:

$$L_i = L_{1i}(\underline{y}_i | \underline{x}_i) L_{2i}(\underline{x}_i)$$

then the overall likelihood is given by

$$L \equiv L_1 L_2 = \prod_{i=1}^n L_{1i}(\underline{y}_i | \underline{x}_i) \prod_{i=1}^n L_{2i}(\underline{x}_i)$$

Note that all the information concerning ρ_{xy} is contained in L_1 . Hence, it will suffice to maximize L_1 rather than L . We can write L_{1i} as a conditional regression likelihood given by

$$L_{1i} = \frac{|\Sigma^*|^{-1/2}}{2\pi} \exp \left[-\frac{1}{2} (\underline{y}_i - \mu_{\underline{y}_i | \underline{x}_i})' \Sigma^{*-1} (\underline{y}_i - \mu_{\underline{y}_i | \underline{x}_i}) \right]$$

where

$$\begin{aligned} \mu_{\underline{y}_i | \underline{x}_i} &= \alpha \mathbf{1} + \beta_1 \underline{x}_i + \beta_2 \underline{v}_i \\ \underline{x}_i &= (x_{i1} \ x_{i2})', \underline{v}_i = (x_{i2} \ x_{i1})', \mathbf{1} = \begin{pmatrix} 1 & 1 \end{pmatrix}' \\ \Sigma^* &\equiv \Sigma_{\underline{y} | \underline{x}} = \sigma_{y|x}^2 \begin{pmatrix} 1 & \rho_{yy|x} \\ \rho_{yy|x} & 1 \end{pmatrix} \end{aligned}$$

Thus, marginally we have the model

$$\begin{aligned} y_{ij} | \underline{x}_i &\sim N \left(\alpha + \beta_1 x_{ij} + \beta_2 x_{i,3-j}, \sigma_{y|x}^2 \right), i=1, \dots, n; j=1, 2 \\ \text{corr}(y_{i1}, y_{i2} | \underline{x}_i) &= \rho_{yy|x} \end{aligned} \quad (1)$$

However, based on the properties of the multivariate normal distribution [1], the conditional distribution of $\underline{y}_i$ given $\underline{x}_i$ is multivariate normal (MVN) with mean and variance given by:

$$\underline{y}_i | \underline{x}_i \sim \text{MVN}(\mu_{\underline{y}_i | \underline{x}_i}, \Sigma_{\underline{y}_i | \underline{x}_i})$$

where

$$\begin{aligned} \mu_{\underline{y}_i | \underline{x}_i} &= \begin{pmatrix} \mu_y \\ \mu_y \end{pmatrix} + \Sigma_{yx} \Sigma_x^{-1} \begin{pmatrix} x_{i1} - \mu_x \\ x_{i2} - \mu_x \end{pmatrix} \\ \Sigma_{\underline{y}_i | \underline{x}_i} &= \Sigma_y - \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy} \end{aligned}$$

In this case, it follows straightforwardly from (1) that

$$\begin{aligned}
E(y_{ij}|\underline{x}_i) &\equiv \mu_{y_i|\underline{x}_i}, \\
&= \mu_y + (\sigma_y/\sigma_x) \left(\frac{\rho_{xy} - \rho_{xy}^* \rho_{xx}}{1 - \rho_{xx}^2} \right) (x_{ij} - \mu_x) + (\sigma_y/\sigma_x) \left(\frac{\rho_{xy}^* - \rho_{xy} \rho_{xx}}{1 - \rho_{xx}^2} \right) (x_{i,3-j} - \mu_x), j=1, 2 \\
V_{jj} &\equiv \sum_{y_i|\underline{x}_i, jj} \equiv \text{Var}(y_{ij}|\underline{x}_i) = \sigma_y^2 \left[1 - \rho_{xy}^2 - \frac{(\rho_{xy} - \rho_{xy} \rho_{xx})^2}{1 - \rho_{xx}^2} \right], j=1, 2 \\
V_{12} &= \sum_{y_i|\underline{x}_i, 12} \equiv \text{Cov}(y_{i1}, y_{i2}|\underline{x}_i) = \sigma_y^2 \left[\rho_{yy} - \frac{(2\rho_{xy}\rho_{xy}^* - \rho_{xy}^2\rho_{xx} - \rho_{xy}^2\rho_{xx})}{1 - \rho_{xx}^2} \right] = \sum_{y_i|\underline{x}_i, 21} \quad (2)
\end{aligned}$$

Therefore, if we equate $E(y_{ij}|\underline{x}_i)$ in equations 1 and 2, we obtain:

$$\beta_1 = \frac{(\rho_{xy} - \rho_{xy}^* \rho_{xx})(\sigma_y/\sigma_x)}{1 - \rho_{xx}^2} \quad (3)$$

$$\beta_2 = \frac{(\rho_{xy}^* - \rho_{xy} \rho_{xx})(\sigma_y/\sigma_x)}{1 - \rho_{xx}^2} \quad (4)$$

Upon combining equations 3 and 4, and solving for ρ_{xy} and ρ_{xy}^* , we obtain:

$$\begin{aligned}
\rho_{xy} &= \frac{\sigma_x}{\sigma_y} (\beta_1 + \rho_{xx} \beta_2) \\
\rho_{xy}^* &= \frac{\sigma_x}{\sigma_y} (\beta_2 + \rho_{xx} \beta_1) \quad (5)
\end{aligned}$$

We can obtain MLE's of σ_x and ρ_{xx} based on $\underline{x}$ from L_2 denoted by $\hat{\sigma}_x$ and $\hat{\rho}_{xx}$ upon using standard mixed effects software (e.g., PROC MIXED of SAS) with a compound symmetry correlation structure. Similarly, we can obtain MLE's of σ_y denoted by $\hat{\sigma}_y$ from the joint likelihood of $\underline{y}_i = (y_{i1}, y_{i2})$ given by $L_3 = \prod_{i=1}^n L_y(\underline{y}_i)$ upon using PROC MIXED based on $\underline{y}$ with a compound symmetry correlation structure.

Thus, we can use equation 5 to obtain MLE's of ρ_{xy} and ρ_{xy}^* given by:

$$\begin{aligned}
\hat{\rho}_{xy, MLE} &= \frac{\hat{\sigma}_x}{\hat{\sigma}_y} \left(\hat{\beta}_1 + \hat{\rho}_{xx} \hat{\beta}_2 \right) \\
\hat{\rho}_{xy, MLE}^* &= \frac{\hat{\sigma}_x}{\hat{\sigma}_y} \left(\hat{\beta}_2 + \hat{\rho}_{xx} \hat{\beta}_1 \right) \quad (6)
\end{aligned}$$

However, an equivalent approach is to express equation 1 directly in terms of ρ_{xy} and ρ_{xy}^* . Specifically, if we substitute equations 3 and 4 into equation 1 and collect terms, we obtain

$$y_{ij}|\underline{x}_i \sim N \left(\alpha + \rho_{xy} s_{ij,2} + \rho_{xy}^* t_{ij,2}, \sigma_{y|\underline{x}}^2 \right), i=1, \dots, n; j=1, 2$$

$$\text{corr}(y_{i1}, y_{i2}|\underline{x}_i) = \rho_{yy|\underline{x}} \quad (7)$$

where

$$s_{ij,2} = \frac{\sigma_y}{\sigma_x} \left(\frac{1}{1-\rho_{xx}^2} \right) (x_{ij} - \rho_{xx} x_{i,3-j})$$

$$t_{ij,2} = \frac{\sigma_y}{\sigma_x} \left(\frac{1}{1-\rho_{xx}^2} \right) (x_{i,3-j} - \rho_{xx} x_{ij})$$

Thus, MLE's of ρ_{xy} and ρ_{xy}^* can be obtained from equation 7 using PROC MIXED with a compound symmetry correlation structure, where σ_y, σ_x and ρ_{xx} are estimated using PROC MIXED in a similar manner as in equation 6.

(b) Clusters of size > 2

Suppose we have data $\underline{z}_i = \begin{pmatrix} \underline{x}_i \\ \underline{y}_i \end{pmatrix}$, where

$\underline{x}_i = (x_{i1}, \dots, x_{ik})'$, $\underline{y}_i = (y_{i1}, \dots, y_{ik})'$ are $(k \times 1)$ vectors and $\underline{z}_i$ is a $2k \times 1$ vector, $i = 1, \dots, n$.

The overall likelihood L is given by:

$$L = \prod_{i=1}^n L_i, \text{ where}$$

$$L_i = L(\underline{z}_i) = L_{1i}(\underline{y}_i|\underline{x}_i) L_{2i}(\underline{x}_i)$$

Since $L_1 \equiv L_1(\rho_{xy}, \rho_{xy}^*, \rho_{yy}, \rho_{xx}, \sigma_y^2, \sigma_x^2)$ and $L_2 \equiv L_2(\rho_{xx}, \sigma_x^2)$, it follows that all the information concerning ρ_{xy} is contained in L_1 . Similar to the case of $k=2$, we can write L_{1i} as a conditional regression likelihood given by:

$$L_{1i} = \frac{|\sum_{y|\underline{x}}|^{-1/2}}{(2\pi)^{k/2}} \exp \left[-\frac{1}{2} (\underline{y}_i - \mu_{\underline{y}_i|\underline{x}_i})' \sum_{y|\underline{x}}^{-1} (\underline{y}_i - \mu_{\underline{y}_i|\underline{x}_i}) \right]$$

where

$$\mu_{\underline{y}_i|\underline{x}_i} = \alpha \mathbf{1}_k + \beta_1 \underline{x}_i + \beta_2 \underline{v}_i$$

$\mathbf{1}_k = (1, \dots, 1)'$ is a $k \times 1$ vector of ones; $\underline{v}_i = (v_{i1}, \dots, v_{ik})'$

and

$$v_{ij} = \sum_{\substack{q=1 \\ q \neq j}}^k x_{iq} \equiv x_{i+,-j}, j=1, \dots, k, i=1, \dots, n$$

$$\Sigma_{y|\underline{x}} = \sigma_{y|\underline{x}}^2 \left[(1 - \rho_{yy|\underline{x}}) I_{k \times k} + \rho_{yy|\underline{x}} J_{k \times k} \right]$$

where $I_{k \times k}$ is the $k \times k$ identity matrix and $J_{k \times k}$ is a $(k \times k)$ matrix of one's.

Thus, marginally, we have the regression model

$$y_{ij} | \underline{x}_i \sim N \left(\alpha + \beta_1 x_{ij} + \beta_2 x_{i+,-j}, \sigma_{y|\underline{x}}^2 \right)$$

$$\text{corr}(y_{ij_1}, y_{ij_2} | \underline{x}_i) = \rho_{yy|\underline{x}}, j_1 \neq j_2 = 1, \dots, k \quad (8)$$

Similar to equation 2, based on the properties of the multivariate normal distribution [1], the conditional distribution of $\underline{y}_i$ given $\underline{x}_i$ is multivariate normal with mean and variance given by

$$\underline{y}_i | \underline{x}_i \sim MVN \left(\mu_{y_i | \underline{x}_i}, \Sigma_{y_i | \underline{x}_i} \right), \text{ where}$$

$$\mu_{y_i | \underline{x}_i} = \mu_y \mathbf{1}_k + \Sigma_{yx} \Sigma_x^{-1} (\underline{x}_i - \mu_x \mathbf{1}_k)$$

$$\Sigma_{y_i | \underline{x}_i} = \Sigma_y - \Sigma_{yx} \Sigma_x^{-1} \Sigma_{xy} \quad (9)$$

It follows straightforwardly from (9) that

$$\mu_{y_i | \underline{x}_i} = \mu_y \mathbf{1}_k + \frac{\sigma_y}{[\sigma_x(1 - \rho_{xx})][1 + (k-1)\rho_{xx}]} \left\{ \begin{aligned} & \left\{ \rho_{xy} [1 + (k-2)\rho_{xx}] - (k-1)\rho_{xy}^* \rho_{xx} \right\} (\underline{x}_i - \mu_x \mathbf{1}_k) \\ & + (\rho_{xy}^* - \rho_{xy} \rho_{xx}) (\underline{y}_i - \mu_y \mathbf{1}_k) \end{aligned} \right\}$$

$$V_{jj} = \Sigma_{y_i | \underline{x}_i, jj} = \text{Var}(y_{ij} | \underline{x}_i) = \sigma_y^2 \left\{ 1 - \rho_{xy}^2 - \frac{(k-1)(\rho_{xy}^* - \rho_{xy} \rho_{xx})^2}{(1 - \rho_{xx})[1 + (k-1)\rho_{xx}]} \right\}, j=1, \dots, k$$

$$V_{j_1, j_2} = \Sigma_{y_i | \underline{x}_i, j_1 j_2} = \text{Cov}(y_{ij_1}, y_{ij_2} | \underline{x}_i) = \sigma_y^2 \left\{ \rho_{yy} - \frac{[2\rho_{xy}\rho_{xy}^* - (k-1)\rho_{xy}^{*2}\rho_{xx} - \rho_{xy}^2\rho_{xx} + (k-2)\rho_{xy}^{*2}]}{(1 - \rho_{xx})[1 + (k-1)\rho_{xx}]} \right\}, j_1 \neq j_2 = 1, \dots, k$$

Note that if $k = 2$, equation 9 reduces to equation 2.

Thus, from equations (8) and (9) we obtain

$$\beta_1 = (\sigma_y / \sigma_x) \left\{ \frac{\rho_{xy} [1 + (k-2)\rho_{xx}] - (k-1)\rho_{xy}^* \rho_{xx}}{[1 + (k-1)\rho_{xx}](1 - \rho_{xx})} \right\} \quad (10)$$

$$\beta_2 = (\sigma_y / \sigma_x) \left\{ \frac{\rho_{xy}^* - \rho_{xy} \rho_{xx}}{[1 + (k-1)\rho_{xx}](1 - \rho_{xx})} \right\}$$

Upon combining equations 8 and 10 and factoring out ρ_{xy} and ρ_{xy}^* , we obtain

$$y_{ij} | \underline{x}_i \sim N \left(\alpha + \rho_{xy} s_{ij,k} + \rho_{xy}^* t_{ij,k}, \sigma_{y|\underline{x}}^2 \right), i=1, \dots, n; j=1, \dots, k \quad (11)$$

where

$$s_{ij,k} = \left(\sigma_y / \sigma_x \right) \frac{\{x_{ij}[1 + (k-2)\rho_{xx}] - x_{i+,-j}\rho_{xx}\}}{[1 + (k-1)\rho_{xx}](1 - \rho_{xx})}$$

$$t_{ij,k} = \left(\sigma_y / \sigma_x \right) \frac{\{x_{i+,-j} - (k-1)x_{ij}\rho_{xx}\}}{[1 + (k-1)\rho_{xx}](1 - \rho_{xx})}$$

and σ_y, σ_x and ρ_{xx} are estimated in a similar manner as in equation 6.

(a) Interval Estimation

We could obtain interval estimates of ρ_{xy} by using $se(\hat{\rho}_{xy})$ from equation 11 and assume asymptotic normality of $\hat{\rho}_{xy}$. However, asymptotic normality is more quickly achieved for estimated correlation coefficients if one uses Fisher's z transformation. If we let $Z_{xy} = 0.5 \ln[(1 + \rho_{xy})/(1 - \rho_{xy})]$ and $\hat{Z}_{xy} = 0.5 \ln[(1 + \hat{\rho}_{xy})/(1 - \hat{\rho}_{xy})]$, then from the delta method we have

$$var(\hat{Z}_{xy}) \cong var(\hat{\rho}_{xy}) / (1 - \hat{\rho}_{xy}^2)^2.$$

Thus, an approximate $100\% \times (1 - \alpha)$ CI for Z_{xy} based on Fisher's z transformation is given by

$$\hat{Z}_{xy} \pm \Phi^{-1}(1 - \alpha/2) se(\hat{\rho}_{xy}) / (1 - \hat{\rho}_{xy}^2) \equiv (Z_{xy,1}, Z_{xy,2})$$

where Φ^{-1} is the inverse normal distribution.

The corresponding $100\% \times (1 - \alpha)$ CI for ρ_{xy} is given by $(\rho_{xy,1}, \rho_{xy,2})$ where

$$\rho_{xy,q} = [\exp(2Z_{xy,q}) - 1] / [\exp(2Z_{xy,q}) + 1], q=1, 2. \quad (12)$$

1. Estimation of Rank Correlation for Clustered Data

We also consider the estimation of rank correlation for clustered data. For this purpose, we define $F_{X_j}(x) \equiv \Pr(X_j \leq x)$ = c.d.f. of X_j , and similarly for $F_{Y_j}(y)$, $j=1, \dots, k$, where X_j and Y_j are continuous random variables. Pearson referred to $F_{X_j}(x)$ and $F_{Y_j}(y)$ as grades of X_j and Y_j within a reference population[4].

Let $H_{X_j}(x) = \Phi^{-1}[F_{X_j}(x)]$, $H_{Y_j}(y) = \Phi^{-1}[F_{Y_j}(y)]$. By definition, if X_j and Y_j are continuous random variables, then H_{X_j} and H_{Y_j} are univariate normal. We also assume that $H_{Z_j} \equiv (H_{X_j}, H_{Y_j})$ is bivariate normal. Finally, we define $P_{iX_j} = F_{X_j}(x_{ij})$ and $P_{iY_j} = F_{Y_j}(y_{ij})$; $i = 1, \dots, n$; $j = 1, \dots, k$.

We wish to estimate $\rho_{s,j} = \text{corr}(P_{iX_j}, P_{iY_j})$ which we define as the underlying rank correlation between X and Y for replicate j . It is customary to estimate $\rho_{s,j}$ by the Spearman rank correlation coefficient $\hat{\rho}_{s,j} = \text{corr}[\text{rank}_j(X_{ij}), \text{rank}_j(Y_{ij})]$ where the ranks are sample ranks of X_{ij} and Y_{ij} , $j = 1, \dots, k$ within a sample of size n . However, we have previously shown that an alternative estimator of $\rho_{s,j}$ exists which is both more efficient than $\hat{\rho}_{s,j}$ and allows one to directly estimate confidence limits for $\rho_{s,j}$ [5]. Specifically, let $H_{iX_j} = \Phi^{-1}(P_{iX_j}) \equiv \text{probit}(P_{iX_j})$, $H_{iY_j} = \Phi^{-1}(P_{iY_j}) \equiv \text{probit}(P_{iY_j})$ and $H_{iZ_j} = (H_{iX_j}, H_{iY_j})$.

We note that the probit transformation is rank-preserving. Therefore, $\rho_{s,j}$ is the same in both the original and probit scale. Let us define $\hat{H}_{iX_j,n} = \Phi^{-1}[\text{rank}_j(X_{ij})/(n+1)]$, $\hat{H}_{iY_j,n} = \Phi^{-1}[\text{rank}_j(Y_{ij})/(n+1)]$ and $\hat{H}_{iZ_j,n} = (\hat{H}_{iX_j,n}, \hat{H}_{iY_j,n})$. In Theorem 2 in the Appendix, we show that if H_{iZ_j} is bivariate normal, then $\hat{H}_{iZ_j,n}$ is asymptotically bivariate normal as $n \rightarrow \infty$. Furthermore, Moran has shown a direct connection between a Pearson and rank correlation for bivariate normally distributed scales [6] whereby

$$\rho_{s,j} = (6/\pi) \sin^{-1}(\rho_{H,j}/2) \quad (13)$$

where $\rho_{H,j} = \text{corr}(H_{iX_j}, H_{iY_j})$. We can estimate $\rho_{H,j}$ by $\hat{\rho}_{H,j} = \text{corr}(\hat{H}_{iX_j,n}, \hat{H}_{iY_j,n})$.

To incorporate clustering, we assume that $\rho_{s,j} = \rho_s$ for all $j = 1, \dots, k$ and use the regression model in equation 11 based on probit scores rather than raw scores to obtain an estimate of $\rho_{H,j} = \text{corr}(H_{iX_j}, H_{iY_j})$ which we assume is the same for all j and denote by $\hat{\rho}_H$. Equation 11 assumes multivariate normality of $\underline{Z}_i = (\underline{X}_i, \underline{Y}_i)$. In Theorem 3 in the Appendix we show that if $(H_1, \dots, H_L)$ is multivariate normal, then $(\hat{H}_{1,n}, \dots, \hat{H}_{L,n})$ is asymptotically multivariate normal as $n \rightarrow \infty$. The corresponding estimator of Spearman rank correlation which is denoted by $\hat{\rho}_{s,a}$ is given by

$$\hat{\rho}_{s,a} = (6/\pi) \sin^{-1}(\hat{\rho}_H/2)$$

Furthermore, a $100\% \times (1-\alpha)$ CI for ρ_s is given by

$$[(6/\pi)\sin^{-1}(\rho_{H,1}/2), (6/\pi)\sin^{-1}(\rho_{H,2}/2)] \quad (14)$$

where $\rho_{H,1}$ and $\rho_{H,2}$ are obtained from equation 12 based on probit scores rather than raw scores.

Similarly, from equation 11 we can also obtain an estimate of the Spearman cross-correlation of (ρ_s^*) given by

$$\hat{\rho}_{s,a}^* = (6/\pi)\sin^{-1}(\hat{\rho}_H^*/2)$$

where

$$\hat{\rho}_H^* = \text{corr}(\hat{H}_{i,X_{j_1}}, \hat{H}_{i,Y_{j_2}}), j_1 \neq j_2 = 1, \dots, k,$$

and obtain a $100\% \times (1-\alpha)$ CI for ρ_s^* based on equations 12 and 13 given by

$$[(6/\pi)\sin^{-1}(\rho_{H,1}^*/2), (6/\pi)\sin^{-1}(\rho_{H,2}^*/2)]$$

2. Estimation of partial correlation for clustered data

It is likely that there is positive correlation between visual field area (VF) and visual acuity (VA). However, it is possible that this is partially due to the effect of age since older RP patients tend to have lower field and acuity measurements than younger RP patients. Also, in some datasets gender differences are found for visual function measures in RP patients [7]. Furthermore, there are different genetic types of RP and level of visual function is known to differ by genetic type [7]. Hence, it would be desirable to estimate the correlation between VF and VA in individual eyes, while controlling for age, gender and genetic type.

(a) Partial Pearson correlation

We assume that there are a set of L cluster-specific covariates denoted by $\underline{u}_i = (u_{i1}, \dots, u_{iL})'$ where $\underline{u}_i$ is an $(L \times 1)$ vector and u_{ij} is the value of the j^{th} cluster-specific covariate for the i^{th} cluster. Let U_i be a $k \times L$ matrix, where $U_{i,jl} = u_{ijl}$, $i = 1, \dots, n$; $j = 1, \dots, k$; $l = 1, \dots, L$. In addition, we assume there are a set of M subunit-specific covariates $\underline{w}_{ij} = (w_{ij1}, \dots, w_{ijM})'$ where $\underline{w}_{ij}$ is an $(M \times 1)$ vector and w_{ijm} is the value of the m^{th} subunit-specific covariate for the j^{th} subunit in the i^{th} cluster. Finally, let W_i be a $k \times M$ matrix where $W_{i,jm} = w_{ijm}$, $i = 1, \dots, n$; $j = 1, \dots, k$; $m = 1, \dots, M$.

We define the partial correlation between X and Y by $\rho_{xy,partial} = corr(x_{ij}, y_{ij} | \underline{u}_i, W_i)$, $i = 1, \dots, n$; $j = 1, \dots, k$. Similarly, the partial cross-correlation between X and Y is defined by $\rho_{xy,partial}^* = corr(x_{ij_1}, y_{ij_2} | \underline{u}_i, W_i)$, $i = 1, \dots, n$; $j_1, j_2 = 1, \dots, k$.

To estimate these correlations, we factor the likelihood of L_i as follows:

$$L_i = L(\underline{z}_i) = L_{Ai}(\underline{z}_i | \underline{u}_i, W_i) L_{Bi}(\underline{u}_i, W_i)$$

Furthermore, we can factor L_{Ai} as follows:

$$L_{Ai}(\underline{z}_i | \underline{u}_i, W_i) = L_{Ai1}(\underline{y}_i | \underline{x}_i, \underline{u}_i, W_i) L_{Ai2}(\underline{x}_i | \underline{u}_i, W_i)$$

All the information concerning $\rho_{xy,partial}$ and $\rho_{xy,partial}^*$ is contained in L_{Ai1} . Hence, it will suffice to maximize $L_{A1} = \prod_{i=1}^n L_{Ai1}$ to obtain the MLE's of $\rho_{xy,partial}$ and $\rho_{xy,partial}^*$. We can write L_{Ai1} as a conditional regression likelihood given by

$$L_{Ai1} = \frac{\left| \sum_{\underline{y} | \underline{x}, U, W} \right|^{-1/2}}{(2\pi)^{k/2}} \exp \left[-\frac{1}{2} \left(\underline{y}_i - \mu_{\underline{y}_i | \underline{x}_i, \underline{u}_i, W_i} \right)' \sum_{\underline{y} | \underline{x}, U, W}^{-1} \left(\underline{y}_i - \mu_{\underline{y}_i | \underline{x}_i, \underline{u}_i, W_i} \right) \right] \quad (15)$$

where

$$\mu_{\underline{y}_i | \underline{x}_i, \underline{u}_i, W_i} = \alpha \mathbf{1}_k + \beta_1 \underline{x}_i + \beta_2 \underline{v}_i + U_i \beta_3 + W_i \beta_4 + W_i^* \beta_5$$

$$W_{i,jm}^* = W_{i+,-j,m} = \sum_{\substack{q=1 \\ q \neq j}}^k w_{iqm}, i=1, \dots, n; j=1, \dots, k; m=1, \dots, M.$$

and

$$\sum_{\underline{y} | \underline{x}, U, W} = \sigma_{\underline{y} | \underline{x}, U, W}^2 \left[\left(1 - \rho_{yy | \underline{x}, U, W} \right) I_{k \times k} + \rho_{yy | \underline{x}, U, W} J_{k \times k} \right]$$

β_3 is an $L \times I$ vector of regression coefficients for cluster-specific covariates and β_4, β_5 are $M \times I$ vectors of regression coefficients for subunit-specific covariates.

To estimate $\rho_{xy,partial}$ and $\rho_{xy,partial}^*$ we

1. Use mixed effects regression models to regress y_i and x_i respectively on U_i , W_i and W_i^* and obtain residuals $\underline{y}_i^*$ and $\underline{x}_i^*$.
2. Obtain estimates of σ_y^* using mixed effects models based on maximizing

$$L_3^* = \prod_{i=1}^n L_{y_i}(\underline{y}_i^*) \quad (16)$$

where $\sigma_y^{2*} = \text{var}(y|U, W) \equiv \text{var}(y_i^*)$

3. Obtain estimates of σ_x^* and ρ_{xx}^* using mixed effects models based on

$$L_2^* = \prod_{i=1}^n L_{2i}(\underline{x}_i^*)$$

where $\sigma_x^{2*} = \text{var}(x|\underline{U}, \underline{W}) \equiv \text{var}(x_i^*)$, $\rho_{xx}^* = \text{corr}(x_{i1}, x_{i2}|\underline{U}, \underline{W})$

4. Obtain covariate-adjusted estimates $s_{ij,k}^*$ and $t_{ij,k}^*$ given by substituting x_{ij}^* , σ_y^* , σ_x^* and ρ_{xx}^* for x_{ij} , σ_y , σ_x and ρ_{xx} in equation 11.
5. Estimate $\rho_{xy, \text{partial}}$ and $\rho_{xy, \text{partial}}^*$ using mixed effects models given by $y_{ij}^* \sim N(\alpha^* + \rho_{xy, \text{partial}} s_{ij,k}^* + \rho_{xy, \text{partial}}^* t_{ij,k}^*, \sigma_{y|x}^{2*})$ with a compound symmetry correlation structure.
6. Obtain interval estimates for $\rho_{xy, \text{partial}}$ and $\rho_{xy, \text{partial}}^*$ using the methods in equation 12.

(b) Partial Spearman rank correlation

Similarly, estimates of $\rho_{s, \text{partial}}$ and $\rho_{s, \text{partial}}^*$ can be obtained by computing sample probits

$\hat{H}_{i, X_j}^*$ and $\hat{H}_{i, Y_j}^*$ using ranks based on x_{ij}^* and y_{ij}^* instead of x_{ij} and y_{ij} and using the methods in equation 13.

5. Unbalanced data

In ophthalmology there often are datasets where some subjects have data available for one eye and other subjects have data available for two eyes. Specifically, in the RP dataset, there were 7 patients with VA and VF available in one eye, but incomplete data in the fellow eye. Five of the patients with incomplete data were missing both VA and VF, while two of the patients were missing VF but not VA, in the incomplete eye. The missing data profile can get more complex if one wants to estimate partial correlation and there is missing covariate data.

In general, since replicates are distinguishable we treat unbalanced data as a type of missing data problem. One possible option is to use all available data, but allow k to vary for different clusters and use mixed effects software methods as provided in equation 11 to

estimate ρ_{xy} . However, as shown in equation 9, the variance-covariance matrix $\Sigma_{y_i^j x_i^j}$ will also depend on k which will not be taken into account using standard software.

If unbalanced data are considered missing data, then the best option is to use multiple imputation methods to analyze the data [8]. For clusters of size 2 using SAS software, this would proceed as follows:

1. use PROC MI of SAS to impute the missing data based on the data vector $h_i = (x_{i1}, y_{i1}, x_{i2}, y_{i2}, \underline{u}_i, \underline{w}_{i1}, \underline{w}_{i2})$ based on M imputations,
2. analyze each imputed data set based on the methods in equation 7,
3. use PROC MIANALYZE to combine results over the M imputations.

This strategy can also be used where the maximum cluster size = k . We used this strategy in both of our examples each of which had small amounts of missing data and had max cluster size = 2 and 3, respectively.

6. Simulation Studies

We have conducted simulation studies to assess the finite sample properties of the estimators of ρ_{xy} both for clusters of 2 as given in Equation 7 and clusters of size > 2 (i.e., 3) as given in Equation 11. We generated $\underline{z}_i$ based on the multivariate normal distribution in Section 2 (separately for $k=2,3$). In addition to the MLE methods described in Equations 7 and 11, we also considered other ad hoc, but frequently used methods. Specifically, we looked at the situation where an investigator has data from 2 eyes available, but only uses a single eye in the analyses (e.g., the right eye, or the left eye) to avoid the correlated data problem. Similarly for $k=3$, where only 1 of 3 distinguishable replicates was used in the analysis. We then estimated ρ_{xy} from the ordinary Pearson product-moment correlation from the n (x,y) pairs and obtained confidence limits based on Fisher's z statistic. We refer to this approach as the *1 subunit method*. In addition, we estimated ρ_{xy} from the ordinary Pearson product-moment correlation based on kn (x,y) pairs and obtained confidence limits based on Fisher's z statistic ignoring the correlated data. We refer to this approach as the *all subunits method*.

We generated multivariate normal data with parameters $\mu = 0$, $\sigma^2 = 1$ and

- a. $\rho_{xy} = \rho_{xy}^* = 0$, $\rho_{xx} = \rho_{yy} = (0.2, 0.5, 0.8)$ to assess type I error.
- b. $(\rho_{xy}, \rho_{xy}^*) = (0.2, 0.1)$, $\rho_{xx} = \rho_{yy} = (0.2, 0.5, 0.8)$ to assess power.
- c. $(\rho_{xy}, \rho_{xy}^*) = (0.5, 0.4)$, $\rho_{xx} = \rho_{yy} = (0.5, 0.8)$ to assess power.

Bias and coverage probability were determined for each method for each of (a) – (c). Separate simulations were performed for $k=2$ and $k=3$. Four thousand simulations were performed for each design in (a) – (c) and each $k=2, 3$. In addition, to assess the adequacy of the model for unbalanced data, we simulated multivariate normal data for $k=3$ subunits per cluster, but randomly deleted the 3rd subunit for 20% of the clusters, thus creating a dataset where 80% of the clusters had 3 subunits and 20% had 2 subunits. The same approach was

used to create a dataset where 95% of the clusters had 3 subunits and 5% of the clusters had 2 subunits. These unbalanced data simulations were performed for each of

- (b) $\rho_{xy} = \rho_{xy}^* = 0, \rho_{xx} = \rho_{yy} = 0.5$ to assess type I error.
- (c) $(\rho_{xy}, \rho_{xy}^*) = (0.2, 0.1), \rho_{xx} = \rho_{yy} = 0.5$ to assess power.

Finally, for each of the designs in (a) – (e) separate simulations were performed for $n=50$ and $n=100$. The results are summarized in Table 1.

For balanced designs, we see that bias is low for all 3 methods. For the MLE method type I error ranges from 0.047 to 0.057 (mean = 0.052), which is acceptable. Similarly, the 1 subunit method type I error ranges from 0.052 to 0.054 (mean = 0.053). However, type I error is not preserved for the all subunits method and ranges from 0.053 to 0.064 (mean = 0.061) for $\rho_{xx} = 0.2$, from 0.077 to 0.120 (mean = 0.100) for $\rho_{xx} = 0.5$ and from 0.127 to 0.202 (mean = 0.155) for $\rho_{xx} = 0.8$.

Coverage probability for the MLE method ranges from 93.6% to 95.9% (mean = 95.0%), which is acceptable. Similarly, coverage probability for the 1 subunit method ranges from 94.6% to 95.0% (mean = 94.8%), which is also acceptable. However, for the all subunits method coverage ranges from 93.7% to 99.3% (mean = 94.8%) when $\rho_{xx} = 0.2$, from 88.0% to 98.7% (mean = 93.4%) when $\rho_{xx} = 0.5$ and from 79.4% to 97.0% (mean = 84.3%) when $\rho_{xx} = 0.8$. In general, the high type I error when $\rho_{xx} = 0.5$ or 0.8 is due to an inappropriately low *se* and a corresponding low coverage probability.

Power for the MLE method is always greater than for the 1 subunit method, sometimes dramatically so, especially when $\rho_{xx} = 0.2$. For example, when $\rho_{xy} = 0.2, \rho_{xy}^* = 0.1$ and $n = 100$, power is 79.9 for the MLE method and 53.1% for the 1 subunit method when $k=2$ and 92.0% for the MLE method and 53.1% for the 1 subunit method for $k=3$. Overall, the MLE method has appropriate type I error and coverage probability and improved power vs. the 1 subunit method. The all subunits method has inappropriate type I error especially when $\rho_{xx} = 0.5$. Thus, overall the MLE method is the preferred method among the 3 approaches.

For unbalanced designs, with 5% of subjects missing a 3rd replicate, bias is low and coverage probability ranges from 94.7% to 95.4%. With 20% of subjects missing a 3rd replicate, there was slight negative bias both for $n=50$ and to a lesser extent $n=100$. Power in the 5% missing data design was 81.8% for $n=100$ and 52.6% for $n=50$, virtually the same as with no missing data ($k=3, n=100, 81.7\%; n=50, 52.4\%$). Power in the 20% missing data design was noticeably reduced ($n=100, 76.9\%; n=50, 40.7\%$). Overall, missing data methods are an effective approach to analyzing clustered data with unbalanced designs.

An assumption of the estimation procedure for rank correlation is that the distribution of both X and Y are continuous. We also performed simulation studies to assess the adequacy of estimation of rank correlation for clustered data based on (13,14) when the distribution of X and/or Y is discrete. For this purpose, we simulated (X,Y) scores from a bivariate normal distribution, computed the sample probits of X and Y scores, but then subdivided the sample probit scores into quintiles and replaced the individual sample probit scores by the median

sample probit score within a quintile. We then obtained point and interval estimates of the rank correlation based on (13) and (14). We considered the following parameter combinations $(\rho_{xy}, \rho_{xy}^*) = (0, 0), (0.2, 0.1)$; $n = (50, 100)$, $(\rho_{xx} = \rho_{yy} = 0.5)$, $k = 3, 4, 1000$ simulations were conducted for each parameter combination. The results are given in Appendix Table 1.

We see that when $(\rho_{xy}, \rho_{xy}^*) = (0, 0)$, the type I error ranges from 0.054 to 0.055. Over the 4 parameter combinations the coverage probability ranges from 94.6% to 95.9%. Power ranges from 52.4% when $n=50$ and 76.8% when $n=100$, which is slightly lower than for continuous distributions with the same parameter combinations $(\rho_{xy}, \rho_{xy}^*) = (0.2, 0.1)$, $(\rho_{xx} = \rho_{yy} = 0.5)$, $k = 3$, power = 81.7%, $n=100$; 52.4%, $n=50$). Overall, the approach in (13) and (14) for estimation of rank correlation seems adequate for discrete distributions based on these designs.

Another assumption of the procedure for estimation of rank correlation for clustered data is that the distribution of (H_X, H_Y) is bivariate normal. By definition H_X and H_Y are univariate $N(0,1)$ random variables, but the assumption of bivariate normality may not hold. To assess the sensitivity of our rank correlation estimation procedures to this assumption we generated (H_X, H_Y) from a Kimeldorf-Sampson (KS) bivariate copula [9] of the form:

$$F(a, b, \delta) = \Pr(H_X \leq a \text{ and } H_Y \leq b) \\ = \left\{ [\Phi(a)]^{-\delta} + [\Phi(b)]^{-\delta} - 1 \right\}^{-1/\delta}, \quad \delta \geq 0. \quad (17)$$

Note that the univariate margins of F are $N(0,1)$ random variables. Also, if $\delta = 0$, then H_X and H_Y are independent, i.e., $F(a, b, 0) = \Phi(a) \Phi(b)$.

To simulate data from this family we first obtained the conditional distribution of $H_Y | H_X$ from (17) given by:

$$F_{H_Y | H_X}(b|a, \delta) = \left\{ [\Phi(a)]^{-\delta} + [\Phi(b)]^{-\delta} - 1 \right\}^{-\frac{1}{\delta}-1} [\Phi(a)]^{-\delta-1} \quad (18)$$

We then generated two iid $U(0,1)$ random variables (U_1, U_2) to correspond to $\Phi(a)$ and $F_{H_Y | H_X}(b|a, \delta)$ respectively, and solved for a and b as follows:

$$a = \Phi^{-1}(U_1) \\ b = \Phi^{-1} \left[\left\{ 1 - U_1^{-\delta} \left[1 - U_2^{-\delta/(1+\delta)} \right] \right\}^{-1/\delta} \right] \quad (19)$$

To simulate clustered data from this family, we created a clustered KS copula by first generating (a, b) as in (19). We then generated iid $(e_{a_q}, e_{b_q}), q = 1, 2$ with univariate $N(0, g)$ margins from this copula based on

$$e_{a_q} = g^{\frac{1}{2}} \Phi^{-1}(U_{1q})$$

$$e_{b_q} = g^{1/2} \Phi^{-1} \left[\left\{ 1 - U_{1q}^{-\delta} \left[1 - U_{2q}^{-\delta/(1+\delta)} \right] \right\}^{-1/\delta} \right], q=1, 2 \quad (20)$$

where (U_{1q}, U_{2q}) are iid $U(0,1)$, (U_{11}, U_{21}) is independent of (U_{12}, U_{22}) and (U_{1q}, U_{2q}) , $q=1,2$ are independent of (U_1, U_2) . Finally, we created clustered data using

$$x_q = (a + e_{a_q}) / \sqrt{1+g}, y_q = (b + e_{b_q}) / \sqrt{1+g}, q=1, 2. \quad (21)$$

Note that based on (17–21) x_q and y_q are univariate $N(0,1)$, but $H(x_q, y_q)$ is not bivariate normal,

$$\rho_{xy} = \text{corr}(x_q, y_q), q=1, 2; \rho_{xy}^* = \text{corr}(x_{q1}, y_{q2}), q_1 \neq q_2=1, 2, \text{ and } \rho_{xy} \neq \rho_{xy}^*$$

To determine the value of δ we refer to Joe [10], page 147 who shows by numerical integration that with copulas from the family in (17) for $\delta = 1.06$, the Spearman rank correlation = 0.5. However, the clustered KS copula in (21) is a summation of two random variables from this family and does not correspond to the distribution of (H_X, H_Y) in (17). Hence, we simulated 100,000 random vectors (X_1, Y_1, X_2, Y_2) from (21) for $g = 1/2$ and $g = 1$, respectively, and empirically estimated $\rho_{xy,s}$ and $\rho_{xy,s}^*$ from

$$\hat{\rho}_{xy,s} = \{ \text{corr}[\text{rank}(X_1), \text{rank}(Y_1)] + \text{corr}[\text{rank}(X_2), \text{rank}(Y_2)] \} / 2$$

$$\hat{\rho}_{xy,s}^* = \{ \text{corr}[\text{rank}(X_1), \text{rank}(Y_2)] + \text{corr}[\text{rank}(X_2), \text{rank}(Y_1)] \} / 2$$

Using $\delta = 1.06$, this yielded $(\hat{\rho}_{xy,s}, \hat{\rho}_{xy,s}^*) = (0.497, 0.328)$ for $g = 1/2$ and $(0.497, 0.246)$ for $g = 1$.

We then assumed these were the true values of $\rho_{xy,s}$ and $\rho_{xy,s}^*$ and performed the estimation procedure in (13) and (14) to obtain point and interval estimates of $\rho_{xy,s}$ based on 4,000 samples of size 50 and 100, respectively, thus obtaining empirical estimates of bias and coverage probability. The results are provided in Appendix Table 2.

We see that the median estimated value is slightly lower than the true value for both $\rho_{xy,s}$ and $\rho_{xy,s}^*$, although the absolute bias is small (0.011). The coverage probability ranges from 94.1% to 96.6% for $g = 1/2$ and 91.4% to 96.7% for $g = 1$ (overall mean = 94.8%) which is acceptable. In summary, the estimation procedure for rank correlation in (13) and (14) performs well for probit scores (H_X, H_Y) whose distribution is not bivariate normal, but instead comes from the clustered KS copula in (21).

5. Examples

(a) Estimating the correlation between visual field and visual acuity in RP patients

We have a sample of 220 RP patients who were enrolled in a clinical trial assessing the effects of docosahexaenoic acid (DHA) supplementation vs. placebo on the rate of decline of visual function over a 4-year period [11]. For this analysis we assess the correlation between Humphrey 30-2 visual field (VF) and ETDRS visual acuity (VA) in individual eyes at the initial (screening) visit. Two hundred thirteen of the patients had VF and VA available in both eyes; five patients had (VA, VF) available in a single eye, but neither available in the fellow eye; two patients had (VA, VF) available in one eye and VA, but not VF available in the fellow eye. Covariate values (i.e., age, gender and genetic type) were complete for all subjects. We used multiple imputation methods as described in Section 5 to impute the missing VF and VA values and perform the analyses described in sections 2, 3 and 4.

We first provide descriptive statistics of VA and VF in Table 2. We see that the mean ETDRS score (VA) = 51.5 ± 0.6 (mean $\pm$ se) letters while the mean Humphrey 30-2 visual field (VF) = 848.9 ± 31.7 db. For reference purposes, a person with 20/20 vision has an ETDRS score of 85 letters; an ETDRS score of 50 letters corresponds roughly to a visual acuity of 20/100 [12]. A normal Humphrey 30-2 visual field is about 2500 db. We see that VF but not VA is lower for older subjects and lower for males vs. females. Furthermore, both VA and VF vary substantially by genetic type with the best function among dominant patients and the worst function among sex-linked patients. Other genetic types had intermediate levels of function. We now provide estimates of the Pearson correlation between VF and VA for individual eyes using the eye as the unit of analysis based on the model in equation 7. The results are given in Table 3. We see that the estimated Pearson correlation between VF(y) and VA(x) in the same eye = 0.212 (95% CI=0.088, 0.329), $p < 0.001$, while the Pearson cross-correlation (correlation between VF in the right eye and VA in the left eye, and conversely) = 0.177 (95% CI = 0.054, 0.295), $p = 0.004$. We also computed a partial Pearson correlation, adjusting for age, gender, and genetic type (person-specific covariates) and right vs. left eye (eye-specific covariate) based on equation 16. The partial Pearson correlation = 0.195 and the partial Pearson cross-correlation = 0.156 which are lower than the crude correlations, although still statistically significant ($p = 0.002$, 0.012, respectively).

Because both VF and VA are not normally distributed, we also computed the corresponding crude and partial rank correlations based on equation 13. The rank correlation coefficients are slightly larger than the corresponding Pearson correlations (ordinary rank correlation = 0.229, $p < 0.001$; partial rank correlation = 0.221, $p < 0.001$; ordinary rank cross-correlation = 0.200, $p < 0.001$; partial rank cross-correlation = 0.189, $p = 0.001$).

Note that for both Pearson and rank correlation, the cross-correlations (ρ_{xy}^* , $\rho_{xy,s}^*$) are almost as large as the corresponding inter-class correlations ($\rho_{xy}\rho_{xy,s}$) indicating likely efficiency gains in estimation of ρ_{xy} and $\rho_{xy,s}$ by $\hat{\rho}_{xy,MLE}$ as presented in equation 7 vs. the ordinary Pearson product-moment correlation given by $\hat{\rho}_{xy,Pearson}$ since $\sigma_{y|x}^2$ will be reduced if t_{ij2} is included as a predictor of y_{ij} in equation 7.

To check the assumption of bivariate normality, we note that if H_X is univariate normal, then (H_X, H_Y) will be bivariate normal if $H_Y/H_X \sim N(\alpha + \beta H_X, \sigma^2)$. To test this assumption we (1) ran a linear regression of $\hat{H}_{i,Y_{j,n}}$ on $\hat{H}_{i,X_{j,n}}$ (2) obtained studentized residuals from the regression and (3) performed the Shapiro-Wilk test of normality on the distribution of studentized residuals (available in PROC UNIVARIATE of SAS). This was done separately for right ($j=1$) and left ($j=2$) eyes, where $Y = VF$ and $X = VA$. The Shapiro-Wilk test statistic was $W = 0.98$, $p > 0.05$ for right eyes and $W = 0.98$, $p > 0.05$ for left eyes indicating that the assumption of bivariate normality is appropriate for these data.

(b) Estimating the correlation between biomarkers of phenacetin and aspirin intake in a study of analgesic abuse among Swiss women

The Swiss Analgesic Study started in 1967/1968 [13]. There were 1244 Swiss women participating in the study. The goal was to evaluate the association between the use of phenacetin-containing analgesics and the prevalence and incidence of kidney disorders. NAPAP (N-acetyl-P-aminophenol) is a urinary metabolite associated with recent use of phenacetin-containing analgesics. NAPAP was initially measured in a urine sample at the baseline clinic visit. There were two additional urine samples collected at home on 2 separate days within 1 week of the baseline clinic visit. One issue is that some analgesics contain both phenacetin and aspirin and it is important to determine the correlation between biomarkers that reflect recent phenacetin and aspirin intake. Salicylate concentrations in urine may be increased after recent intake of aspirin-containing analgesics, but the amount of increase depends on the pH level in urine with alkaline urine ($pH > 7.0$) resulting in increased secretion in urine [14]. Urinary salicylates may also increase after ingestion of other substances (e.g., cranberry juice) [15].

For this analysis, concentrations of NAPAP and salicylates were determined from the same 3 urine specimens for each woman. NAPAP values were represented as continuous variables and expressed in units of optical density (o.d.), but its distribution is highly skewed and non-normal. Urinary salicylates (mg %) were measured in four categories (0-19 mg% / 20-49 mg% / 50-99 mg% / 100+ mg%) and converted to a continuous scale by using the median value within each category (10 mg% / 35 mg% / 75 mg% / 100 mg%). Since the distributions of both NAPAP and salicylates were non-normal, the rank correlation is a natural measure of association. Descriptive data for this study are provided in Table 4.

There were 90 women (7.2%) who had missing data for at least one NAPAP replicate and/or one salicylate replicate. We used multiple imputation methods with 5 imputations to impute the missing data (using PROC MI of SAS) and to perform the analyses in equation 11 (using PROC MIANALYZE of SAS).

We see that mean NAPAP was higher for women age 40 or older vs. women < 40 years of age, while mean salicylates were comparable in the two age groups. Mean NAPAP was higher for both light (57.0kg) and heavier (73.0kg) women and was much higher for women who currently smoke 10 cigarettes/day vs. non-current smokers. The trends for salicylates were in the same direction. Both NAPAP and salicylates were somewhat higher at the clinic visit than at the two home urines.

The correlation analyses based on equation 11 are reported in Table 5. The intraclass correlation (ICC) was higher for NAPAP ($\rho_{yy} = 0.611$) than for salicylates ($\rho_{xx} = 0.402$). The Pearson correlation between NAPAP and salicylates was 0.391 ($p < 0.001$) when measured at the same visit (ρ_{xy}) and 0.248 when measured at different visits (ρ_{xy}^*). After adjusting for age, weight, smoking and location of urine collection, the estimated Pearson correlation between NAPAP and salicylates remained about the same when measured at the same visit (i.e., $\rho_{xy} = 0.390$) and when measured at different visits (i.e., $\rho_{xy}^* = 0.246$).

Since the distribution of NAPAP was highly skewed and the distribution of salicylates represented in 4 broad ordered categories, it was reasonable to also consider rank correlation as a measure of association. The estimated rank correlation was 0.290 when NAPAP and salicylates were measured at the same visit ($\rho_{xy,s}$) and 0.197 when measured at different visits ($\rho_{xy,s}^*$). After adjusting for age, weight, smoking and location of urine collection, the rank correlations decreased slightly when measured at the same visit ($\rho_{xy,s} = 0.288$) and at different visits ($\rho_{xy,s}^* = 0.194$).

This example shows that there is a considerable difference between $\hat{\rho}_{xy}$ vs. $\hat{\rho}_{xy}^*$ and similarly between $\hat{\rho}_{xy,s}$ vs. $\hat{\rho}_{xy,s}^*$ indicating that NAPAP and salicylate values at different visits are best modeled as distinguishable replicates.

We also performed the Shapiro-Wilk test of normality for the distribution of studentized residuals of $\hat{H}_{i,Y_{j,n}}$ given $\hat{H}_{i,X_{j,n}}$ as described in the RP example. The Shapiro-Wilk test statistic was $W = 0.99$, $p < 0.05$ at the clinic visit, $W = 0.99$, $p < 0.05$ at the 1st home urine, and $W = 0.99$, $p < 0.05$ at the 2nd home urine. The reference value for the W test statistic under normality is $W = 1.0$. Thus, there was only a mild departure from the null value of W and the significance of the W test is due to the large sample size.

6. Discussion

In this paper we have presented methods for maximum likelihood estimation of Pearson correlation in the clustered data setting. The methods are applicable to clustered data based on distinguishable replicates and can be used both for clusters of size 2 (Equation 7) and clusters of size > 2 (Equation 11). They are based on a population-average type of model. A simpler approach to this problem from a population-average perspective is based on the regression model

$$y_{ij} = \alpha + \beta x_{ij} + e_{ij}, i = 1, \dots, n; j = 1, \dots, k \quad (22)$$

where

$$e_{ij} \sim MVN(\mathbf{0}, \Sigma)$$

and

$$\Sigma = \sigma_{y|x}^2 \left[(1 - \rho_{yy|x}) I_{k \times k} + \rho_{yy|x} J_{k \times k} \right]$$

The corresponding estimator of ρ_{xy} denoted by $\hat{\rho}_{xy, standard}$ is given by:

$$\hat{\rho}_{xy, standard} = \hat{\beta} \hat{\sigma}_x / \hat{\sigma}_y \quad (23)$$

The model in (22) and the corresponding estimator in (23) are actually a special case of Equation 11 when $\rho_{xy}^* = \rho_{xy} \rho_{xx}$. In words, $\rho_{xy}^* = \rho_{xy} \rho_{xx}$ if the correlation between y_{ij1} and x_{ij2} is completely mediated by the correlation between y_{ij} and x_{ij} and between x_{ij1} and x_{ij2} . In general, this will not always be the case and the estimator in (23) is not the MLE of ρ_{xy} under the model in (8).

Bland and Altman [16] have also considered a population-average approach to estimation of ρ_{xy} . With their approach, a pseudo-data set is created where pairs $(\bar{x}_i, \bar{y}_i)$ are repeated k_i times and $k_i = \#$ replicates for the i^{th} subject. Their estimator can be expressed in the form:

$$\hat{\rho}_{xy, BA} = \sum_{i=1}^n k_i (\bar{x}_i - \bar{\bar{x}})(\bar{y}_i - \bar{\bar{y}}) / \left\{ \left[\sum_{i=1}^n k_i (\bar{x}_i - \bar{\bar{x}})^2 \right] \left[\sum_{i=1}^n k_i (\bar{y}_i - \bar{\bar{y}})^2 \right] \right\}^{1/2}$$

where

$$\bar{\bar{x}} = \sum_{i=1}^n k_i \bar{x}_i / K, \bar{\bar{y}} = \sum_{i=1}^n k_i \bar{y}_i / K, K = \sum_{i=1}^n k_i$$

The ordinary Pearson correlation is then computed based on the K observations in the pseudo-data set and is tested for significance using a t distribution with $n - 2$ df as the reference distribution.

Lorenz, Datta and Harkema [17] also consider a population-average approach, but based on observations from individual cluster-members rather than cluster averages. Their estimator takes the form:

$$\hat{\rho}_{xy, LDH} = \sum_{i=1}^n \left[\sum_{j=1}^{k_i} (x_{ij} - \bar{\bar{x}})(y_{ij} - \bar{\bar{y}}) / k_i \right] / \left\{ \left[\sum_{i=1}^n \sum_{j=1}^{k_i} (x_{ij} - \bar{\bar{x}})^2 / k_i \right] \left[\sum_{i=1}^n \sum_{j=1}^{k_i} (y_{ij} - \bar{\bar{y}})^2 / k_i \right] \right\}^{1/2}$$

Large sample asymptotic normality of $\hat{\rho}_{xy, LDH}$ is assumed and significance testing is based on $z_{LDH} = \hat{\rho}_{xy, LDH} / [\text{var}(\hat{\rho}_{xy, LDH})^{1/2}]$ and compared with a $N(0,1)$ distribution. The delta

method is used to estimate $\text{var}(\hat{\rho}_{xy,LDH})$. The LDH approach is also applied to estimation of Kendall's τ for clustered data.

Both the BA and LDH approaches assume that replicates are indistinguishable and allow for a variable number of replicates per cluster. With the BA estimator, the i^{th} cluster-mean is weighted k_i times. With the LDH estimator each cluster gets equal weight regardless of cluster-size. If either $\rho_{xx} > 0$ or $\rho_{yy} > 0$, the optimal weighting for the i th cluster will be between 1 and k_i .

The estimators in equations 7 and 11 of this paper assume replicates are distinguishable and use multiple imputation methods to estimate missing (x, y) replicate scores based on (x, y) scores for other cluster members as well as values of other covariates and performs the analysis based on n_{IMP} imputed data sets using Rubin's rules. The estimators are maximum likelihood, can be implemented with standard software, are extended to the rank correlation setting and allow one to control for other covariates.

The distinguishing feature of ML estimation is that for clustered data the likelihood is based on

$$L = \prod_{i=1}^n L_i,$$

where $L_i = \prod_{j=1}^{k_i} L_{1i}(y_{ij}|x_{ij}) L_{2i}(x_{ij})$ and $L_1 = \prod_{i=1}^n L_{1i}$ is maximized to estimate ρ_{xy} .

In the LDH approach the likelihood is based on

$$L_1^* = \prod_{i=1}^n \prod_{j=1}^{k_i} L(y_{ij}|x_{ij})$$

Under L_1^* , y_{ij} is expressed as a function of x_{ij} , while under L_1 , y_{ij} is expressed as a function of both x_{ij} and $x_{i+,-j}$. L_1^* is a special case of L_1 when $\rho_{xy}^* = \rho_{xy}\rho_{xx}$. L_1^* is the likelihood in the i.i.d. situation when $\rho_{xx} = \rho_{yy} = 0$ [1], but not in the clustered data situation.

Another approach is to consider a subject-specific type of model [18]. Under this approach, we consider the random intercept model

$$y_{ij} = \mu + q_i + \beta x_{ij} + e_{ij}, \quad i=1, \dots, n; j=1, \dots, k_i \quad (24)$$

where

$$q_i \sim N(0, \sigma_q^2), e_{ij} \sim N(0, \sigma^2)$$

Under this model, the estimator β has a subject-specific interpretation (i.e., are changes in X between 2 replicates for an individual correlated with corresponding changes in Y). For example, do changes in visual field between 2 eyes of a subject correspond to changes in visual acuity. Bland and Altman [18] consider a slight variation of (24) where q is a fixed effect instead of a random effect.

Finally, we have extended our results to the estimation of Spearman rank correlation for clustered data. To our knowledge, this is the first time this has appeared in the literature. We also extend the Pearson and rank correlation results to the setting of partial correlation, where one can control for both cluster-specific and subunit-specific covariates. For rank correlation this is different from the partial Spearman correlation available in some statistical packages where one calculates covariate-adjusted residuals for both Y and X based on the raw data and then computes the Spearman correlation based on the residuals. With our approach, probit scores are first computed for X and Y variables and residuals are computed by regressing X and Y probit scores respectively, on covariates. Rank correlations are then obtained based on Spearman correlations between the probit residuals. This should provide a more nonparametric estimate of partial Spearman correlation.

The relationship between Pearson and Spearman correlation given in (13), assumes that $H_Z = (H_X, H_Y)$ is bivariate normal. If X and/or Y are categorical, such as for salicylates in the second example, then this assumption will be violated. However, if we assume that there is a latent continuous scale but the continuous scale is divided into categories and the actual rank of an observation on the latent continuous scale is replaced by the average rank within a category, then this assumption may be more plausible. Also, in the simulation studies where we replaced the actual rank on the continuous scale by the average rank within a category (see Appendix Table 1), the type I error and coverage were preserved.

We also assessed the validity of the estimation procedure in (13) and (14) when H_X and H_Y are each univariate normal, but (H_X, H_Y) is not bivariate normal. The clustered KS copula in (21) was introduced for this purpose. Simulation studies indicate that bias and coverage probability associated with the estimators of $\rho_{xy,s}$ and $\rho_{xy,s}^*$ in (13) and (14) are acceptable in this setting for $n = 50$ (see Appendix Table 2). We also proposed a diagnostic procedure to test the assumption of bivariate normality of replicate probit scores which was applied and satisfied for both of the examples in this paper.

Acknowledgments

This work was supported by the National Eye Institute, R01 EY022445. The authors would also like to acknowledge the programming support of Marion McPhee.

References

1. Anderson, TW. An introduction to multivariate statistical analysis. Wiley; New York: 1958.

2. Jennrich RI, Schluchter MD. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*. 1986; 42:805–820. [PubMed: 3814725]
3. Group ETDRS. Early Therapy Diabetic Retinopathy Study (ETDRS) manual of operations. University of Maryland; City: 1985. Early Therapy Diabetic Retinopathy Study (ETDRS) manual of operations.
4. Pearson, K. *Mathematical Contributions to the Theory of Evolution. On further methods of determining correlation*. Cambridge University Press; Cambridge, UK: 1907.
5. Rosner B, Glynn RJ. Interval estimation for rank correlation coefficients based on the probit transformation with extension to measurement error correction of correlated ranked data. *Stat Med*. 2007; 26:633–646. [PubMed: 16596580]
6. Moran PA. Rank correlation and product-moment correlation. *Biometrika*. 1948; 35:203–206. [PubMed: 18867425]
7. Berson EL, Sandberg MA, Rosner B, Birch DG, Hanson AH. Natural course of retinitis pigmentosa over a three-year interval. *Am J Ophthalmol*. 1985; 99:240–251. [PubMed: 3976802]
8. Rubin, DB. *Multiple imputation for non-response in surveys*. John Wiley and Sons; New York: 1987.
9. Kimeldorf G, Sampson AR. Uniform representation of bivariate distributions. *Comm Statist*. 1975; 4:617–627.
10. Joe, H. *Multivariate models and dependence concepts*. Chapman & Hall; London ; New York: 1997.
11. Berson EL, Rosner B, Sandberg MA, Weigel-DiFranco C, Moser A, Brockhurst RJ, Hayes KC, Johnson CA, Anderson EJ, Gaudio AR, Willett WC, Schaefer EJ. Clinical Trial of docosahexaenoic acid in patients with retinitis pigmentosa receiving vitamin A treatment. *Archives of Ophthalmology*. 2004; 122:1297–1305. [PubMed: 15364708]
12. Gregori NZ, Feuer W, Rosenfeld PJ. Novel method for analyzing snellen visual acuity measurements. *Retina*. 2010; 30:1046–1050. [PubMed: 20559157]
13. Dubach UC, Levy PS, Muller A. Relationships Between Regular Analgesic Intake and Urorenal Disorders in a Working Female Population of Switzerland I. Initial Results (1968). *American Journal of Epidemiology*. 1971; 93:425–434. [PubMed: 5562715]
14. Proudfoot AT, Krenzelok EP, Brent J, Vale JA. Does urine alkalization increase salicylate elimination? If so, why? *Toxicol Rev*. 2003; 22:129–136. [PubMed: 15181662]
15. Duthie GG, Kyle JA, Jenkinson AM, Duthie SJ, Baxter GJ, Paterson JR. Increased salicylate concentrations in urine of human volunteers after consumption of cranberry juice. *J Agric Food Chem*. 2005; 53:2897–2900. [PubMed: 15826036]
16. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part 2--Correlation between subjects. *British Medical Journal*. 1995; 310:633. [PubMed: 7703752]
17. Lorenz DJ, Datta S, Harkema SJ. Marginal association measures for clustered data. *Stat Med*. 2011; 30:3181–3191. [PubMed: 21953204]
18. Bland JM, Altman DG. Calculating correlation coefficients with repeated observations: Part 1--Correlation within subjects. *British Medical Journal*. 1995; 310:446. [PubMed: 7873953]
19. Lehmann, EL. *Nonparametrics: Statistical Methods Based on Ranks*. Springer; New York: 2006. (revised edn)

Appendix – Proofs

Theorem 1

Let X be a continuous random variable and let $\{x_1, \dots, x_n\}$ be an i.i.d. sample from X . Define

$\hat{H}_{x,n} = \Phi^{-1} \left[R_{x,n}^{(x)} / (n+1) \right]$, where Φ^{-1} is the inverse normal distribution,

$R_{x,n}^{(x)} = 1 + \sum_{i=1}^n U(x - x_i)$, and $U(a) = 1$ if $a > 0$, $= 0$ otherwise. If $H_X = \Phi^{-1}[F_X(x)]$ where F_X is the c.d.f. of X , then $\hat{H}_{x,n}$ converges in probability to H_X as $n \rightarrow \infty$.

Proof

From (A.110), p. 346 of Lehmann [19] if $E(\hat{H}_{x,n} - H_x)^2 \rightarrow 0$ as $n \rightarrow \infty$, then $\hat{H}_{x,n}$ converges in probability to H_x as $n \rightarrow \infty$.

Let $P_x = \Phi(H_x) = F_X(x)$ and $\hat{P}_{x,n} = \Phi(\hat{H}_{x,n}) = R_{x,n}^{(x)} / (n+1)$

We have that

$$E(\hat{H}_{x,n} - H_x)^2 = [E(\hat{H}_{x,n}) - H_x]^2 + \text{var}(\hat{H}_{x,n}) \quad (\text{A1})$$

To assess the 1st component of (A1) we use a Taylor series expansion of $\hat{H}_{x,n}$ about H_x , yielding

$$\hat{H}_{x,n} = H_x + \sqrt{2\pi} \exp(H_x^2/2) (\hat{P}_{x,n} - P_x) + \pi H_x \exp(H_x^2) (\hat{P}_{x,n} - P_x)^2 + \sum_{r=3}^{\infty} a_r (\hat{P}_{x,n} - P_x)^r \quad (\text{A2})$$

where $|a_r| < \infty$, $r \geq 3$. if we take expectations of both sides of (A2) we obtain:

$$E(\hat{H}_{x,n}) = H_x + \sqrt{2\pi} \exp(H_x^2/2) E(\hat{P}_{x,n} - P_x) + \pi H_x \exp(H_x^2) E(\hat{P}_{x,n} - P_x)^2 + \sum_{r=3}^{\infty} a_r E(\hat{P}_{x,n} - P_x)^r \quad (\text{A3})$$

Since $E(\hat{P}_{x,n} - P_x)^q \rightarrow 0$ as $n \rightarrow \infty$ for $q \geq 1$ it follows that $E(\hat{H}_{x,n}) - H_x \rightarrow 0$ as $n \rightarrow \infty$

To assess the 2nd component of (A1) we use the delta method, whereby

$$\begin{aligned} \text{var}(\hat{H}_{x,n}) &= 2\pi \exp(H_x^2) \text{var}(\hat{P}_{x,n}) + O(1/n) \\ &= 2\pi \exp(H_x^2) n P_x (1 - P_x) / (n+1)^2 + O(1/n) \rightarrow 0 \text{ as } n \rightarrow \infty \end{aligned} \quad (\text{A4})$$

Thus, based on (A1), (A3) and (A4), $E(\hat{H}_{x,n} - H_x)^2 \rightarrow 0$ as $n \rightarrow \infty$, whereby $\hat{H}_{x,n}$ converges in probability to H_x as $n \rightarrow \infty$.

Corollary 1

Let $\hat{H}_{X,n}$ be defined as in Theorem 1, $H_X = \{\hat{H}_X, x \in X\}$ and $\hat{H}_{X,n} = \{\hat{H}_{X,n}, x \in X\}$. $\hat{H}_{X,n}$ converges in law to a $N(0,1)$ distribution as $n \rightarrow \infty$.

Proof

We have that

$$\lim_{n \rightarrow \infty} \Pr(\hat{H}_{X,n} \leq z) = \Pr(H_X \leq z) = \Phi(z).$$

Hence, $\hat{H}_{X,n}$ is asymptotically normal as $n \rightarrow \infty$.

Theorem 2

Let $\hat{H}_{X,n}$ and $\hat{H}_{Y,n}$ be defined as in Corollary 1. Let $H_Z = (H_X, H_Y)$, $\rho_{XY} = \text{cov}(H_X, H_Y)$ and $\hat{H}_{Z,n} = (\hat{H}_{X,n}, \hat{H}_{Y,n})$. If H_Z is bivariate normal, then $\hat{H}_{Z,n}$ is asymptotically bivariate normal as $n \rightarrow \infty$.

Proof

From Theorem 2.6.2, p. 37 of Anderson [1], H_Z is bivariate normal if and only if $aH_X + bH_Y$ is univariate normal for all a, b , where $\min(|a|, |b|) > 0$. We have from Corollary 1 that $a\hat{H}_{X,n} + b\hat{H}_{Y,n}$ converges in law to $aH_X + bH_Y$ as $n \rightarrow \infty$. Thus, since $\text{var}(H_X) = \text{var}(H_Y) = 1$ and $\text{cov}(H_X, H_Y) = \rho_{XY}$ then

$$\lim_{n \rightarrow \infty} \Pr(a\hat{H}_{X,n} + b\hat{H}_{Y,n} \leq z) = \Phi\left(z / \sqrt{a^2 + b^2 + 2ab\rho_{XY}}\right)$$

Therefore, $a\hat{H}_{X,n} + b\hat{H}_{Y,n}$ is asymptotically univariate normal for all a, b . It follows that $H_{Z,n}$ is asymptotically bivariate normal as $n \rightarrow \infty$.

Theorem 3

Let $X_1, \dots, X_L$ be continuous random variables and define $H_{X_l}(x) = \Phi^{-1}[F_{X_l}(x)]$, $l=1, \dots, L$. Let $\rho_{l_1 l_2} = \text{cov}(H_{X_{l_1}}, H_{X_{l_2}})$, $l_1, l_2 = 1, \dots, L$. Define $\hat{H}_{X_{l,n}}(x) = \Phi^{-1}[R_{X_{l,n}}(x)/(n+1)]$, where $R_{X_{l,n}}(x) = 1 + \sum_{i=1}^n U(x - x_{il})$ and x_{il} = value of X_l for the i^{th} subject, $i = 1, \dots, n$; $l = 1, \dots, L$.

If $H_X \equiv (H_{X_1}, \dots, H_{X_L})$ is multivariate normal, then $\hat{H}_{X,n} \equiv (\hat{H}_{X_{1,n}}, \dots, \hat{H}_{X_{L,n}})$ will be asymptotically multivariate normal as $n \rightarrow \infty$.

Proof

From Theorem 2.6.2, p. 37 of Anderson [1] H_X will be multivariate normal if and only if

$Z = \sum_{l=1}^L a_l H_{X_l}$ is univariate normal for all $\underline{a} = (a_1, \dots, a_L)$ where $\min(|a_l|, l=1, \dots, L) > 0$. We

have from Corollary 1 that $\hat{Z}_n = \sum_{l=1}^L a_l \hat{H}_{X_l, n}$ converges in law to Z as $n \rightarrow \infty$. Thus, since $\text{var}(H_{X_l}) = 1$, $l = 1, \dots, L$, and $\text{cov}(H_{X_{l_1}}, H_{X_{l_2}}) = \rho_{l_1 l_2}$, $l_1, l_2 = 1, \dots, L$, then

$\lim_{n \rightarrow \infty} \Pr(\hat{Z}_n \leq z) = \Phi\left(z / \sqrt{\sum_{l=1}^L a_l^2 + \sum_{l_1 \neq l_2=1}^L a_{l_1} a_{l_2} \rho_{l_1 l_2}}\right)$. Therefore, $\hat{Z}_n$ is asymptotically univariate normal for all z where $\min(|a_l|, l = 1, \dots, L) > 0$. It follows that $\hat{H}_{X_n}$ is asymptotically multivariate normal as $n \rightarrow \infty$.

Appendix Table 1

Simulation Results for Rank Correlation Based on Discrete Data, MLE method

k	ρ_{xy}	ρ_{xy}^*	$\rho_{xx}(\rho_{yy})$	n	$\hat{\rho}_{xy}$	Type I error	coverage probability (%)	power
3	0	0	0.5	50	0.028	0.055	94.7	---
	0	0	0.5	100	0.010	0.054	94.8	---
	0.2	0.1	0.5	50	0.194	---	95.9	52.4
	0.2	0.1	0.5	100	0.184	---	94.6	76.8

Computer runs:

:/proj/stross/stros0a/rankcorrelation/simulation.3rep.4000.n100.5.probit.sas 9/15/16

:/proj/stross/stros0a/rankcorrelation/simulation2.3rep.4000.n100.5.probit.sas 9/15/16

:/proj/stross/stros0a/rankcorrelation/simulation.3rep.4000.n50.5.probit.sas 9/15/16

:/proj/stross/stros0a/rankcorrelation/simulation2.3rep.4000.n50.5.probit.sas 9/15/16

Appendix Table 2

Bias and coverage probability of $(\hat{\rho}_{xy, s}, \hat{\rho}_{xy, s}^*)$ based on (13) and (14) when the data are generated from the clustered KS* copula in (21)

g	n	parameter	true value**	median estimated value ⁺	bias	coverage probability (%)
0.5	50	$\rho_{xy, s}$	0.497	0.488	-0.009	96.6
		$\rho_{xy, s}^*$	0.328	0.327	-0.001	94.4
	100	$\rho_{xy, s}$	0.497	0.491	-0.006	96.6
		$\rho_{xy, s}^*$	0.328	0.325	-0.003	94.1
	50	$\rho_{xy, s}$	0.497	0.486	-0.011	96.4
		$\rho_{xy, s}^*$	0.246	0.246	0	92.1
1.0	100	$\rho_{xy, s}$	0.497	0.491	-0.006	96.7
		$\rho_{xy, s}^*$	0.246	0.244	-0.002	91.4

* Kimeldorf-Sampson

** based on simulating 100,000 random vectors (X_1, Y_1, X_2, Y_2) from the clustered KS copula in (21) and estimating $\rho_{xy, s}$ from mean [Spearman correlation (X_1, Y_1) + Spearman correlation (X_2, Y_2)] and $\rho_{xy, s}^*$ from mean [Spearman correlation (X_1, Y_2) + Spearman correlation (X_2, Y_1)].

⁺ based on 4,000 simulated datasets consisting of $n(X_1, Y_1, X_2, Y_2)$ vectors from the clustered KS copula in (21).

Computer runs:

```
:/proj/stross/stros0a/rankcorrelation/simulation.new.n100000.sas 9/29/16  
:/proj/stross/stros0a/rankcorrelation/simulation.new2.n100000.sas 10/13/16  
:/proj/stross/stros0a/rankcorrelation/simulation.new.4000.n50.sas 10/13/16  
:/proj/stross/stros0a/rankcorrelation/simulation.new.4000.n100.sas 10/13/16  
:/proj/stross/stros0a/rankcorrelation/simulation.new2.4000.n50.sas 10/14/16  
:/proj/stross/stros0a/rankcorrelation/simulation.new2.4000.n100.sas 10/14/16
```

Figure 1.
Distribution of Humphrey 30-2 field (decibels) for right (OD) and left (OS) eyes among 220 retinitis pigmentosa (RP) patients

Figure 2.
Distribution of ETDRS visual acuity (number of letters correctly read) for right (OD) and left (OS) eyes among 220 retinitis pigmentosa (RP) patients

Figure 3 (a)

Figure 3 (b)

Figure 3.

Figure 3a Cross-classification of Humphrey 30-2 field (decibels) for right (OD) and left (OS) eyes among 220 retinitis pigmentosa (RP) patients

Figure 3b Cross-classification of ETDRS visual acuity (number of letters correctly read) for right (OD) and left (OS) eyes among 220 retinitis pigmentosa (RP) patients

Table 1

Simulation Results

k	ρ_{xy}	ρ_{xy}^*	$\rho_{xx}(\rho_{yy})$	method	n=50			n=100				
					$\hat{\rho}_{xy}$	Type I error	coverage prob.	power	$\hat{\rho}_{xy}$	Type I error	coverage prob.	power
Balanced Designs												
2	0	0	0.2	MLE	0.001	0.052	94.9	---	0.001	0.047	95.3	---
				1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---
				all subunits	0.000	0.061	94.0	---	0.001	0.053	99.3	---
				MLE	0.001	0.056	94.4	---	0.001	0.050	95.2	---
				1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---
				all subunits	0.001	0.091	90.9	---	0.001	0.077	98.7	---
	0.2	0.1	0.2	MLE	0.001	0.057	94.6	---	0.002	0.050	95.2	---
				1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---
				all subunits	0.000	0.133	86.7	---	0.002	0.127	97.0	---
				MLE	0.200	---	95.0	49.3	0.200	---	95.3	79.9
				1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
				all subunits	0.199	---	94.4	51.4	0.200	---	94.8	81.1
3	0.5	0.4	0.5	MLE	0.201	---	94.7	43.3	0.200	---	95.3	72.0
				1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
				all subunits	0.200	---	91.2	52.5	0.201	---	92.1	78.9
				MLE	0.201	---	94.9	33.2	0.202	---	95.2	61.1
				1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
				all subunits	0.201	---	86.4	52.4	0.202	---	87.1	76.7
	0.5	0.4	0.5	MLE	0.499	---	94.6	99.8	0.500	---	94.6	100.0
				1 subunit	0.497	---	94.9	96.4	0.500	---	95.0	100.0
				all subunits	0.496	---	92.6	99.9	0.498	---	92.9	100.0
				MLE	0.500	---	95.5	99.0	0.502	---	95.3	100.0
				1 subunit	0.497	---	94.9	96.4	0.500	---	95.0	100.0
				all subunits	0.498	---	87.1	99.7	0.500	---	87.6	100.0
0	0	0.2	MLE	0.001	0.052	95.0	---	0.000	0.056	94.5	---	
			1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---	

k	ρ_{xy}	ρ_{xy}^*	$\rho_{xx}(\rho_{yy})$	method	$n=50$				$n=100$			
					$\hat{\rho}_{xy}$	Type I error	coverage prob.	power	$\hat{\rho}_{xy}$	Type I error	coverage prob.	power
2	0.5			all subunits	0.001	0.060	94.0	---	0.000	0.064	93.7	---
				MLE	0.003	0.052	95.2	---	0.001	0.057	94.6	---
				1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---
				all subunits	0.003	0.113	88.7	---	0.001	0.120	88.0	---
	0.8			MLE	0.004	0.048	95.7	---	0.001	0.052	95.1	---
				1 subunit	-0.001	0.054	94.6	---	0.002	0.052	94.8	---
				all subunits	0.004	0.197	80.3	---	0.002	0.202	79.8	---
				MLE	0.201	---	95.1	66.6	0.200	---	94.4	92.0
	0.2	0.1		1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
				all subunits	0.199	---	94.2	68.6	0.199	---	93.7	92.9
				MLE	0.203	---	95.2	52.4	0.202	---	94.6	81.7
				1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
3	0.5			all subunits	0.202	---	88.9	66.5	0.201	---	88.1	89.7
				MLE	0.205	---	95.4	32.4	0.202	---	95.1	63.7
				1 subunit	0.198	---	94.6	28.3	0.201	---	94.7	53.1
				all subunits	0.204	---	79.8	34.9	0.201	---	79.4	84.8
	0.8	0.4		MLE	0.500	---	94.6	100.0	0.500	---	93.6	100.0
				1 subunit	0.497	---	94.9	96.4	0.500	---	95.0	100.0
				all subunits	0.496	---	90.3	100.0	0.498	---	90.3	100.0
				MLE	0.503	---	95.9	99.3	0.501	---	95.2	100.0
	0.2	0.1		1 subunit	0.497	---	94.9	96.4	0.500	---	95.0	100.0
				all subunits	0.500	---	80.4	100.0	0.499	---	79.5	100.0
				MLE	0.002	0.055	95.2	---	0.000	0.058	94.7	---
				MLE	0.202	---	95.4	52.6	0.201	---	94.7	81.8
	0.5	0.1		MLE	-0.019	0.049	95.5	---	-0.008	0.056	94.8	---
				MLE	0.182	---	95.2	40.7	0.193	---	94.5	76.9

Unbalanced Designs

* 4000 simulations were run for each set of parameter combinations

Table 2

Descriptive Statistics for the RP Data Set (220 patients)*

Variable	Category	VA**			VF***		
		mean (se)	N _{eyes} (N _{eyes,obs} [†])	(95% CI)	mean (se)	N _{eyes} (N _{eyes,obs} [†])	(95% CI)
age (yrs)	overall	51.5 (0.6)	440 (435)	(50.3, 52.8)	848.9 (31.7)	440 (433)	(786.7, 911.1)
	< 35	50.3 (1.0)	166 (166)	(48.3, 52.3)	934.3 (51.2)	166 (165)	(833.9, 1034.6)
	35	52.3 (1.0)	274 (269)	(50.7, 53.8)	797.2 (39.9)	274 (268)	(719.0, 875.4)
gender	M	51.0 (0.9)	218 (217)	(49.2, 52.7)	798.4 (44.9)	218 (216)	(710.4, 886.3)
	F	52.1 (0.9)	222 (218)	(50.3, 53.8)	898.5 (44.5)	222 (217)	(811.3, 985.7)
genetic type [‡]	DOM	54.4 (1.3)	102 (100)	(51.9, 56.9)	1005.6 (65.1)	102 (99)	(878.0, 1133.2)
	AR	51.3 (1.7)	60 (60)	(48.1, 54.6)	849.6 (84.9)	60 (59)	(683.3, 1015.5)
	XL	44.1 (2.3)	30 (30)	(39.5, 48.7)	670.6 (119.9)	30 (30)	(435.5, 905.7)
	ISO	52.2 (0.9)	214 (211)	(50.5, 53.9)	812.1 (45.0)	214 (211)	(724.0, 900.2)
	UND	45.4 (2.2)	34 (34)	(41.0, 49.7)	766.3 (112.7)	34 (34)	(545.4, 987.1)
	eye						
	right eye (OD)	51.5 (0.6)	220 (219)	(50.3, 52.7)	873.0 (31.7)	220 (217)	(810.4, 935.5)
	left eye (OS)	51.6 (0.7)	220 (216)	(50.1, 53.0)	824.8 (32.6)	220 (216)	(760.6, 889.0)

* VA is available for 219 patients for the right eye (OD) and 216 patients for the left eye (OS); VF is available for 217 patients for the right eye (OD) and 216 patients for the left eye (OS); age, gender and genetic type are available for all 220 patients; missing data are accounted for using multiple imputation.

** VA = visual acuity (ETDRS letter score)

*** VF = Humphrey visual field (db)

[‡]DOM = dominant; AR = recessive; XL = sex-linked; ISO = isolate; UND = undetermined

[†]N_{eyes,obs} = number of eyes observed for the characteristic; N_{eyes} = number of eyes observed + number of eyes imputed for the characteristic computer run:

./proj/stross/stross0a/rankcorrelation/correlation.proc.mi.age.sas 11/24/15

Author Manuscript Author Manuscript Author Manuscript Author Manuscript
:/proj/stross/stross0a/rankcorrelation/correlation.ranked.ns.sas 11/7/16
:/proj/stross/stross0a/rankcorrelation/correlation.ranked.proc.mi.overall 12/1/15
:/proj/stross/stross0a/rankcorrelation/correlation.ranked.proc.mi.sexddm 11/30/15

Table 3

Correlation Estimates for the RP Data Set (220 patients, 440 eyes) *

Type of correlation	Parameter	Crude				Adjusted			
		Point estimate	95% CI	P-value	N _{eyes} (N _{eyes,obs} **)	Point estimate	95% CI	P-value	N _{eyes} (N _{eyes,obs} **)
Pearson	ρ_{xy}	0.212	(0.088, 0.329)	<0.001	440 (433)	0.195	(0.071, 0.312)	0.002	440 (433)
	ρ_{xy}^*	0.177	(0.054, 0.295)	0.004	440 (433)	0.156	(0.033, 0.274)	0.012	440 (433)
	ρ_{xx}	0.787				0.766			
	σ_x^2	100.2				91.1			
	ρ_{yy}	0.944				0.943			
Rank	σ_y^2	227, 330				201, 710			
	ρ_{xys}	0.229	(0.110, 0.341)	<0.001	440 (433)	0.221	(0.103, 0.333)	<0.001	440 (433)
	$\rho_{xy,s}^*$	0.200	(0.082, 0.312)	<0.001	440 (433)	0.189	(0.072, 0.301)	0.001	440 (433)
	ρ_{xxs}	0.819				0.807			
	$\sigma_{x,s}^2$	0.946				0.891			
	ρ_{yys}	0.909				0.900			
	$\sigma_{y,s}^2$	0.962				0.884			

* VA is available for 219 patients for the right eye (OD) and 216 patients for the left eye (OS); VF is available for 217 patients for the right eye (OD) and 216 patients for the left eye (OS); age, gender and genetic type are available for all 220 patients; missing data are accounted for using multiple imputation.

** N_{eyes,obs} = number of eyes where both VA and VF are observed; N_{eyes} = number of eyes observed + number of eyes imputed for the characteristic
computer run:

```
:/proj/stross/stross0a/rankcorrelation/correlation.ranked,proc.mi3a.sas 11/23/15
```

```
:/proj/stross/stross0a/rankcorrelation/correlation.ranked,proc.mi.residual.probit.sas 12/04/15
```

```
:/proj/stross/stross0a/rankcorrelation/correlation.ranked,proc.mi.residual.sas 12/04/15
```

Table 4

Descriptive Statistics for the Swiss Study Data Set, (1244 women) *

	NAPAP (o.d.)			Salicylates (mg%)		
	mean(se)	N _{vis} (N _{vis,obs})**	(95% CI)	mean(se)	N _{vis} (N _{vis,obs})**	(95% CI)
overall	0.162 (0.008)	3732 (3680)	(0.147, 0.178)	15.0 (0.4)	3732 (3600)	(14.2, 15.7)
age						
<40	0.146 (0.013)	1311 (1292)	(0.120, 0.172)	15.2 (0.7)	1311 (1265)	(13.9, 16.5)
40	0.171 (0.010)	2421 (2388)	(0.152, 0.191)	14.8 (0.5)	2421 (2335)	(13.9, 15.8)
weight (kg)						
57.0	0.202 (0.016)	912 (898)	(0.171, 0.233)	16.4 (0.8)	912 (880)	(14.9, 18.0)
57.1 – 63.9	0.134 (0.017)	840 (832)	(0.102, 0.167)	13.2 (0.8)	840 (813)	(11.6, 14.8)
64.0 – 72.9	0.142 (0.014)	1113 (1098)	(0.114, 0.170)	14.6 (0.7)	1113 (1075)	(13.2, 16.1)
73.0	0.174 (0.016)	867 (851)	(0.142, 0.206)	15.6 (0.8)	867 (832)	(14.0, 17.2)
Number of cigarettes currently smoked per day						
0	0.134 (0.010)	2295 (2266)	(0.117, 0.156)	14.5 (0.5)	2295 (2222)	(13.5, 15.5)
1 – 9	0.157 (0.017)	759 (745)	(0.123, 0.191)	15.4 (0.9)	759 (725)	(13.7, 17.1)
10	0.258 (0.018)	678 (669)	(0.222, 0.294)	16.1 (0.9)	678 (653)	(14.4, 17.9)
location of urine collection						
clinic visit	0.175 (0.010)	1244 (1244)	(0.155, 0.194)	15.6 (0.5)	1244 (1202)	(14.5, 16.6)
1 st home urine	0.153 (0.009)	1244 (1217)	(0.136, 0.170)	14.6 (0.5)	1244 (1191)	(13.7, 15.6)
2 nd home urine	0.160 (0.009)	1244 (1219)	(0.142, 0.177)	14.7 (0.5)	1244 (1207)	(13.7, 15.7)

* NAPAP is available for 1244 women at the clinic visit, 1217 women at the 1st home urine and 1219 women for the 2nd home urine; Salicylates are available for 1202 women at the clinic visit, 1191 women at the 1st home urine and 1207 women at the 2nd home urine; age, weight and number of cigarettes currently smoked are available for all 1244 women; missing data are accounted for using multiple imputation.

** N_{vis,obs} = number of visits where the characteristic was observed; N_{vis} = number of visits where the characteristic was observed + number of visits where the characteristic was imputed.

:/proj/stross/stros0a/rankcorrelation/swiss.proc.mi.age.sas 01/22/16

:/proj/stross/stros0a/rankcorrelation/swiss.proc.mi.overall.sas 01/22/16

:/proj/stross/stros0a/rankcorrelation/swiss.proc.mi.weight.smoke.sas 02/02/16

:/proj/stross/stros0a/rankcorrelation/swiss.age.wgt.smk.ms.sas 11/7/16

Table 5

Correlation Estimates for the Swiss Study Data Set (1244 women)*

Type of correlation	Parameter	Crude [†]				Adjusted [‡]			
		Point estimate	95% CI	P-value	N _{vis} (N _{vis,obs}) ^{**}	Point estimate	95% CI	P-value	N _{vis} (N _{vis,obs}) ^{**}
Pearson	ρ_{xy}	0.391	(0.355, 0.426)	<0.001	3732 (3597)	0.390	(0.354, 0.425)	<0.001	3732 (3597)
	ρ_{xy}^*	0.248	(0.213, 0.282)	<0.001	3732 (3597)	0.246	(0.212, 0.280)	<0.001	3732 (3597)
	ρ_{xx}	0.402				0.401			
	σ_x^2	313.2				312.5			
	ρ_{yy}	0.611				0.602			
	σ_y^2	0.105				0.102			
Rank	$\rho_{xy,s}$	0.290	(0.250, 0.328)	<0.001	3732 (3597)	0.288	(0.247, 0.328)	<0.001	3732 (3597)
	$\rho_{xy,s}^*$	0.197	(0.162, 0.231)	<0.001	3732 (3597)	0.194	(0.159, 0.229)	<0.001	3732 (3597)
	$\rho_{xx,s}$	0.329				0.329			
	$\sigma_{x,s}^2$	0.426				0.426			
	$\rho_{yy,s}$	0.538				0.528			
	$\sigma_{y,s}^2$	0.952				0.932			

* NAPAP is available for 1244 women at the clinic visit, 1217 women at the 1st home urine and 1219 women for the 2nd home urine; Salicylates are available for 1202 women at the clinic visit, 1191 women at the 1st home urine and 1207 women at the 2nd home urine; age, weight and number of cigarettes currently smoked are available for all 1244 women; missing data are accounted for using multiple imputation.

** N_{vis,obs} = number of visits where both urinary NAPAP and salicylates were available; N_{vis} = number of visits where both urinary NAPAP and salicylates were available + number of visits where either was imputed

[†] y=NAPAP, x=salicylates

[‡] adjusted for age, weight, smoking and location of urine collection computer run:

:/proj/stross/stross0a/rankcorrelation/swiss.proc.mi.probit.residual3.sas 2/4/16

:/proj/stross/stross0a/rankcorrelation/swiss.proc.mi.residual3.sas 2/2/16

:/proj/stross/stross0a/rankcorrelation/swiss.proc.mi.probit.sas 1/21/16

:/proj/stross/stross0a/rankcorrelation/swiss.proc.mi2.sas 1/21/16

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript