

# The Intraclass Covariance Matrix

Gregory Carey<sup>1,2</sup>

Received 03 March 2005—Final 10 May 2005

Introduced by C.R. Rao in 1945, the intraclass covariance matrix has seen little use in behavioral genetic research, despite the fact that it was developed to deal with family data. Here, I reintroduce this matrix, and outline its estimation and basic properties for data sets on pairs of relatives. The intraclass covariance matrix is appropriate whenever the research design or mathematical model treats the ordering of the members of a pair as random. Because the matrix has only one estimate of a population variance and covariance, both the observed matrix and the residual matrix from a fitted model are easy to inspect visually; there is no need to mentally average homologous statistics. Fitting a model to the intraclass matrix also gives the same log likelihood, likelihood-ratio (LR)  $\chi^2$ , and parameter estimates as fitting that model to the raw data. A major advantage of the intraclass matrix is that only two factors influence the LR  $\chi^2$ —the sampling error in estimating population parameters and the discrepancy between the model and the observed statistics. The more frequently used interclass covariance matrix adds a third factor to the  $\chi^2$ —sampling error of homologous statistics. Because of this, the degrees of freedom for fitting models to an intraclass matrix differ from fitting that model to an interclass matrix. Future research is needed to establish differences in power—if any—between the interclass and the intraclass matrix.

**KEY WORDS:** Twins; Sib-Pairs; linear models; covariance matrix; correlation matrix; interclass covariance; intraclass covariance.

In current behavioral genetic research on twins or sib-pairs, scores for the second member of a pair are concatenated to those for the first member and then a covariance matrix is calculated. This covariance matrix is an *interclass* matrix. In some cases, the assignment of members to “twin 1” versus “twin 2” (or sib 1 versus sib 2) is dictated by the research design. For example, if pairs are ascertained because of a single affected individual, then the proband may be designated as “twin 1” and the cotwin as “twin 2.” Another example is the covariance matrix for opposite-sex DZ twins where females, say, are twin 1 and males are twin 2.

In many other cases, however, the assignment of members to twin 1 or twin 2 is random or arbitrary with respect to the design or the models fitted to the data. In this case, the vector of means for twin 1 and the vector of means for twin 2 have the same expectation, the within-person covariance matrix for twin 1 and the within-person covariance matrix for twin 2 have the same expectation, and the cross-twin covariance matrix is symmetric. Here, the twins are in an *intraclass* relationship. Note that the choice of interclass or intraclass relationships is not a property of the twins *per se* but depends on the research design and the models fitted to data. In one analysis of the same data set, opposite-sex DZ pairs could be treated as being in an interclass relationship while the next analysis could deal with them as an intraclass relationship.

When the design or model implies an intraclass relationship, then the researcher may calculate the intraclass covariance matrix and fit models to that

<sup>1</sup> Department of Psychology and Institute for Behavioral Genetics, University of Colorado, Boulder, CO, 80309-0345, USA.

<sup>2</sup> To whom correspondence should be addressed at Department of Psychology and Institute for Behavioral Genetics, University of Colorado, Boulder. Tel: +1-303-492-1658; Fax: +1-303-492-2967; e-mail: gregory.carey@colorado.edu

matrix. This strategy, however, is seldom employed in behavioral genetic research. The purpose of this paper is to explain the intraclass covariance matrix and outline several of its basic properties.<sup>1</sup>

### INTRACLASS STATISTICS

Let  $k$  denote the number of phenotypes. Here, I develop the intraclass model for a single type of twins (e.g., MZ or DZ pairs) and later extend it to multiple types. In the intraclass model, the expected mean for any phenotype for the first member of the pair is the same as that for the second member. Letting  $\mu_i$  denote the mean for the  $i$ th phenotype, the vector of expected values for an intraclass relationship may be written as

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_k, \mu_1, \mu_2, \dots, \mu_k)^t.$$

Because of the intraclass relationship, the within-person covariance matrix,  $\mathbf{W}$ , for the first twin will equal that of the second twin, and the cross-twin covariance matrix,  $\mathbf{T}$ , will be symmetric. Hence, the expected intraclass covariance matrix,  $\boldsymbol{\Sigma}$ , will be partitioned as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{W} & \mathbf{T} \\ \mathbf{T} & \mathbf{W} \end{pmatrix}. \quad (1)$$

Note that an interclass matrix contains what I shall call homologous statistics—two estimates of the same population variance or covariance. For example, the variance of the first phenotype for twin 1 and the variance for the first phenotype for twin 2 are homologous statistics. Usually, these are different numbers, so visual inspection of the matrix requires the mental averaging of two numbers. The intraclass matrix, on the other hand, has one and only one estimate for a population variance or covariance. Hence, there is no need for mental averaging in this matrix.

### INTRACLASS ESTIMATORS

Let the vector of scores for the  $i$ th twin pair be represented as

$$\mathbf{x}_i = (X_{11i}, X_{12i}, \dots, X_{1ki}, X_{21i}, X_{22i}, \dots, X_{2ki})^t,$$

where the first subscript denotes the first or second member of the pair and the second subscript denotes

the phenotype. If the phenotypes are sampled from a multivariate normal, then the log likelihood (less a constant) for a sample of  $N$  pairs is

$$\log(L) = -\frac{N}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}).$$

Tedious algebraic reduction for the first derivatives of this function with respect to  $\boldsymbol{\mu}$  reveal that the maximum likelihood (ML) estimate of the expected value for the  $k$ th phenotype is

$$\hat{\boldsymbol{\mu}}_k = \frac{\sum_{i=1}^N X_{1ki} + X_{2ki}}{2N}.$$

This is simply the sample mean treating the raw scores of twin 1 and twin 2 as if they were independent observations.

Let  $\mathbf{D}$  denote the matrix of the deviations of raw scores from their means and let  $\mathbf{d}_i^t$  be the  $i$ th row vector from this matrix. Then the log likelihood of the scores corrected for the mean is

$$\begin{aligned} \log(L) &= -\frac{N}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^N \mathbf{d}_i^t \boldsymbol{\Sigma}^{-1} \mathbf{d}_i \\ &= -\frac{N}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \text{trace}(\mathbf{D}^t \mathbf{D} \boldsymbol{\Sigma}^{-1}). \end{aligned} \quad (2)$$

Using the derivatives for matrix elements given by Mulaik (1972), it can be shown that the maximum likelihood estimate for a variance is

$$\hat{\sigma}_k^2 = \frac{\sum_{i=1}^N D_{1ki}^2 + D_{2ki}^2}{2N},$$

the maximum likelihood estimate of a within-individual covariance is

$$\hat{W}_{jk} = \frac{\sum_{i=1}^N D_{1ji} D_{1ki} + D_{2ji} D_{2ki}}{2N},$$

and the maximum likelihood of a cross-twin covariance is

$$\hat{T}_{jk} = \frac{\sum_{i=1}^N D_{1ji} D_{2ki} + D_{1ki} D_{2ji}}{2N}.$$

While it is possible to use numerical methods to obtain estimates for the means and the covariance matrix, a more efficient strategy is to create a data set by double entering the twin pairs (first enter twin 1's scores followed by twin 2's scores, then enter twins 2's

<sup>1</sup> I cannot detail all of the properties of intraclass relationships here, especially the extension to variable sibship size. For further details, the inquisitive reader should consult C.R. Rao (1945), who is credited with developing the multivariate intraclass covariance matrix, Konishi *et al.*, (1991) and Srivastava *et al.*, (1988).

scores followed by twin 1's scores). The observed means from these data are the maximum likelihood estimators of the means. The corrected sum of squares and cross products matrix (SSCP) matrix from this data set (i.e., corrected for the means) divided by  $2N$  is the maximum likelihood estimate of  $\Sigma$ . Note that when  $N$  is large (as it should be for ML), then the results from computing a covariance matrix using standard statistical packages that divide the SSCP matrix by  $2N - 1$  will be acceptable for all practical purposes.

The estimators point out an important difference between the intraclass and the interclass matrix. For a given data set, there is one and only one intraclass matrix. On the other hand, there is a large number of interclass matrices, each one dependent on which member gets denoted as twin 1 or twin 2.

### FITTING MODELS TO INTRACLASS MATRICES

From now on, let us consider the means as fixed and focus on the covariance matrix. Note that the quantity  $\mathbf{D}^t\mathbf{D}$  may be written as  $N\mathbf{S}$  where  $\mathbf{S}$  is the observed intraclass covariance matrix. Then a base log likelihood can be calculated in which the predicted matrix,  $\Sigma$ , equals the observed intraclass covariance matrix. Using Equation (2), this base log likelihood reduces to

$$\log(L_0) = -\frac{N}{2}(\log|\mathbf{S}| + 2k).$$

Let  $\Sigma_A$  denote the expected covariance matrix under an alternative model. Then the likelihood-ratio (LR)  $\chi^2$  for this model becomes

$$\begin{aligned} \chi^2 &= -2[\log(L_A) - \log(L_0)] \\ &= N(\log|\Sigma_A| + \text{trace}(\mathbf{S}\Sigma_A^{-1}) - \log|\mathbf{S}| - 2k). \end{aligned} \quad (3)$$

Note that this expression is almost identical to the one used for fitting an interclass matrix.<sup>2</sup>

The derivation of Equation (3) leads us to another major property of the intraclass matrix—fitting a model to the intraclass matrix is the same as fitting that model to the raw data. Except

<sup>2</sup> There are two small differences, however, between Equation (3) and the traditional formula used in behavioral genetics. First, in programs like Mx (Neale *et al.*, 2002), the quantity in parentheses in that equation is multiplied by  $(N-1)$ , not by  $N$ . Second, the divisor for the observed covariance matrix is usually  $(N-1)$  while this formula implies that the divisor would be  $N$ . Once again, these differences become trivial with large sample sizes.

for numerical error, one will get the same log likelihood, the same likelihood ratio  $\chi^2$ , and the same parameter estimates. The fit statistics and parameter estimates from fitting a model to the interclass matrix will differ from those derived from the same model fitted to raw data.

Another advantage of using the intraclass covariance matrix comes in assessing the fit of the model by inspection of the residual matrix, i.e., difference between the observed and the predicted covariance matrices. With the intraclass matrix, there is no need to mentally average two homologous residuals to assess fit.

The major difference between fitting models to an interclass and to an intraclass matrix lies in the degrees of freedom (df). With  $k$  phenotypes, the interclass matrix has  $k(2k+1)$  unique observed statistics. From Equation (1), however, the intraclass matrix has  $k(k+1)$  unique observed statistics. Hence, the degrees of freedom for fitting a model with  $p$  parameters to an interclass matrix is  $k(2k+1) - p$ , but the df for fitting the same model to an intraclass matrix is  $k(k+1) - p$ .

The difference in df is easily explained. The LR  $\chi^2$  for the intraclass matrix measures the discrepancy of the model from the data plus some sampling error of the estimators (i.e., the variances and covariances) from their population values. The LR  $\chi^2$  for the interclass matrix also measures this discrepancy but in addition includes the sampling error for homologous statistics. There are  $k^2$  homologous statistics in the interclass matrix, exactly the difference in df between the interclass and intraclass matrices.

Again, there is an advantage to the intraclass matrix. With an interclass matrix, a LR  $\chi^2$  that suggests an acceptable fit could come about from a poor model (which generates a number of LR  $\chi^2$  units) but little sampling error in homologous statistics (which results if few LR  $\chi^2$  units). Similarly, the opposite situation (a decent model but, but dumb luck, significant sampling variation) could give what appears to be a poor fit. The LR  $\chi^2$  for the intraclass matrix does not tap this extraneous source of variance.<sup>3</sup>

There is a major issue, however, that does need to be resolved—power. It is not clear whether there is a difference in power between the two matrices or, if

<sup>3</sup> One can of course, fit a model to an interclass matrix that forces the homologous statistics to be identical and then assess the fit of any subsequent model with respect to this model. That strategy, however, adds another step to the data analysis.

there is a difference, which strategy is more powerful under which situations. The intraclass matrix has one estimate of a population variance or covariance. For homologous statistics, the interclass matrix has two estimates of the same population parameters. Two estimates may indeed be better than one in some instances. Because the purpose of this paper is to reintroduce the intraclass covariance matrix to a behavioral genetics audience, I leave the simulations required to explore power to future research.

### MULTIPLE TYPES

The derivations given above are appropriate for fitting models to one type of twin or sib. The extension to multiple types (e.g., both MZ and DZ twins) is straightforward. Here, one simply calculates the intraclass matrix for each type and fits the model to these matrices using Equation (3) for each type.

Note that it is possible to mix intraclass with interclass matrices. Take, for example, the case of modeling sex differences in twin data. One could use the intraclass covariance matrices for the four types of same-sex twins but the interclass matrix for the

opposite-sex DZ pairs. The only trick is to keep the degrees of freedom straight. With  $k$  phenotypes, there are  $k(k+1)$  unique statistics in each intraclass matrix and  $k(2k+1)$  unique statistics in the DZ-OS matrix. Hence, the df for the LR  $\chi^2$  will equal  $k(6k+5) - p$  where  $p$  is the number of free parameters in the model.

### REFERENCES

- Konishi, S., Khatri, C. G., and Rao, C. R. (1991). Inferences on multivariate measures on interclass and intraclass correlations in familial data. *J. Roy. Stat. Soc. B. (Met)* **53**:649–659.
- Mulaik, S. A. (1972). *The Foundations of Factor Analysis*. New York: McGraw Hill.
- Neale, M. C., Boker, S. M., Xie, G., and Maes, H. (2002). *Mx: Statistical Modeling* (6th ed.). Richmond VA: Virginia Institute for Psychiatric and Behavioral Genetics Virginia Commonwealth University.
- Rao, C. R. (1945). Familial correlations or the multivariate generalizations of the intraclass correlation. *Curr. Sci.* **14**:66–67.
- Srivastava, M. S., Keen, K. J., and Katapa, R. S. (1988). Estimation of interclass and intraclass correlations in multivariate family data. *Biometrics* **44**:141–150.

Edited by Stacey Cherny