Statistica Sinica **12**(2002), 519-535

# A UNIFIED THEORY OF STATISTICAL ANALYSIS AND INFERENCE FOR VARIANCE COMPONENT MODELS FOR DYADIC DATA

Heng Li and Eric Loken

*University of Rochester and The Pennsylvania State University*

*Abstract:* Using the covariance structure induced by the exchangeability of sampling units, a unified approach to the analysis of dyadic data is proposed. Dyadic data, encountered in diallel designs in genetics and other substantive scientific fields, arise when pairs of sampling units are studied. The problem has been addressed independently in a number of different areas of study. This paper argues that dyadic data structures involve the same statistical elements as those of ordinary analysis of variance and multivariate analysis. In addition to a synthesis of the available literature, the article provides a closed form expression of the Gaussian likelihood, the sufficient statistics and their joint distributions, and outlines for EM and ECM algorithms for handling missing data and other complications. The approach is illustrated with an applied example. The objective is to show that the analysis of dyadic data can be developed as a standard statistical method not unlike the analysis of variance, albeit with a multivariate twist. Dyadic data structures can be treated similarly to ordinary factorial structures and have the potential to be more widely used.

*Key words and phrases:* Analysis of variance, Bayesian inference, Bio model, Cockerham-Weir Model, diallel design, EM algorithm, ECM algorithm, exchangeability, maximum likelihood, Social Relations Model.

## 1. Introduction

A common experimental design in genetics is the diallel cross, in which pairs of distinct strains are cross-bred in order to estimate genetic effects. According to Wright (1985, p.307), this design "has probably attracted more attention and been the subject of more theoretical examination and practical application in the past 30 years than any other mating design". ANOVA-style models have been described by many researchers (Yates (1947), Kempthorne (1952), Cockerham and Weir (1977), Simms and Triplett (1996), Zhu and Weir (1996a, b), Husband and Gurney (1998), Lynch and Walsh (1998), Lipow and Wyatt (1999), Xu and Zhu (1999), Motten and Stone (2000)). However, this basic data structure is not unique to genetics. Scientific interest in outcomes based on the pairing

of sampling units arises in psychophysics (Bechtel (1967, 1971)), social psychology (Kenny (1994), Lashley and Bond (1997)), comparative psychology (Lev and Kinder (1957)) and social network analysis (Kraemer and Jacklin (1979), Iacobucci and Wasserman (1987), Wasserman and Faust (1994)).

We use the term "dyadic data" to describe data derived from observations on pairs of sampling units. When all possible pairings are observed, the experimental design is often called either a diallel cross or a round-robin. A traditional analytic approach to dyadic data has been to fit variance component models, with independent developments in the above mentioned fields (Cockerham and Weir (1977), Warner, Kenny and Stoto (1979), Wong (1982), Bond and Lashley (1996)). However, considering the common formal structure in dyadic data with regard to sampling and observation, which does not depend on the content area, it is rather curious that each discipline has separately invented its own approach. Indeed, there has been little recognition of this common formal structure, and of the opportunity to synthesize the separately developed methods within a unified statistical framework. To see the peculiarity of the situation with dyadic data from a historical perspective, imagine that a diverse group of substantive areas developed and used their own equivalent of the two-way analysis of variance, without a central documentation in the statistical literature of the formal procedure, or common knowledge in each area that the same formal procedure was being used in each specific instance. Unfortunately, this is now more or less the case for dyadic data. Our objective is to establish a unified framework for the statistical analysis of dyadic data, and to relate and generalize existing methods within this framework. We begin by providing some background information.

## 2. The Generation of Dyadic Data

### 2.1. Notation

To formally describe the dyadic data structure, we set up a system of subscription to distinguish it from two-way data in an ordinary analysis of variance. We use the symbol $(i, j)$ to indicate an ordered pair formed by distinct sampling units $i$ and $j$ from the same population, and $y_{(i,j)}$ to indicate a single observation on such an ordered pair. Reciprocal dyads are defined as two dyads consisting of the same sampling units in opposite order, i.e., $(i, j)$ and $(j, i)$. In the case of multiple observations on the same dyad, we will use $y_{(i,j)k}$ to denote the $k$th observation on dyad $(i, j)$. The vector of all the data points is denoted by $\mathbf{y}$.

### 2.2. Data generation

The contemplation of the sources of variability for dyadic data can be guided by the accumulated experience with variance component models associated with

ordinary analysis of variance. The dyads can naturally be grouped into reciprocals, and accordingly we can initially decompose the variability into that from within reciprocals, and that from between reciprocals. The within reciprocals and between reciprocals variability can each in turn have two sources, that due to individual units and that due to interaction between individual units. From the above considerations the variance component model for dyadic data should have the form

$$y_{(i,j)} = \mu + g_i + g_j + s_{ij} + d_i - d_j + r_{ij}, \tag{1}$$

where $s_{ij} = s_{ji}$ and $r_{ij} = -r_{ji}$, with $\mu$ being a constant (or random variable), $g$ and $s$ representing between reciprocals variability, $d$ and $t$ representing within reciprocals variability, and all terms except $\mu$ having mean 0. In keeping with the tradition of analysis of variance, we assume that all terms are normally distributed, and the interaction terms are independent of each other and of the individual main effect terms.

Under these assumptions, $y_{(i,j)}$ can be generated as follows: draw two independent $(g, d)$'s from $N(\mathbf{0}, \Sigma_{gd})$ for individuals $i$ and $j$; independently draw $r$ from $N(0, \sigma_r^2)$ and $s$ from $N(0, \sigma_s^2)$; $y_{(i,j)}$ is realized via equation (1). Note that even though we allow for a correlation between $g$ and $d$, variances are still additive in (1). Given $\sigma_s^2$, $\sigma_r^2$, $\sigma_g^2$, $\sigma_d^2$, and $\sigma_{gd}$, $var(y_{(i,j)}) = 2\sigma_g^2 + \sigma_s^2 + 2\sigma_d^2 + \sigma_r^2$, because the two terms involving $\sigma_{gd}$ cancel out. The additivity of variance seems to make it reasonable to call (1) a variance component model. Later we see that (1) can take several equivalent forms.

### 2.3. A justification of the data generation mechanism based on covariance structure

The exchangeability induced by random sampling of units from a population puts constraints on the covariances between observations on dyads. Since the group action on dyads is transitive, the variances of the observations are all the same, denoted by $\sigma^2$. As a result of group invariance, there are at most five different covariances between different dyads, denoted by $\sigma^2\rho_i$, $i = 0, \ldots, 4$. There is a one-to-one correspondence between the variance component parameters in (1) and the above covariance parameters together with $\sigma^2$, if we specify $\mu$ in (1) as a random variable. Therefore, the variance component model (1) is justified by exchangeability in the sense of Dawid (1988). The expressions of the parameters $\sigma^2\rho_i$ $\{i = 0, \ldots, 4\}$ and $\sigma^2$ as functions of those in (1) are given in the third column of Table 1, with $\mu$ specified as a constant, as is the case in the rest of the paper. If we specify $\mu$ in (1) as a random variable, then its variance $\sigma_\mu^2$ would be added to each term in the third column of the table. Translating the covariance structure into the variance component model (1) reveals the structure

of dyadic data. Model (1) indicates that certain dyadic phenomena measured by a continuous variable can potentially be explained by two latent traits of the individual units within the dyads, and two synergistic effects specific to each dyad. Of course, as with the simplest one-way random effects model, the above data generation mechanism, while preserving the covariance pattern required by exchangeability, puts constraints on the covariances more stringent than those guranteeing the nonnegative definiteness of the entire covariance matrix. This implies that the variance component model (1) is not a plausible data generation mechanism when the covariance parameters do not satisfy those constraints, similar to the one-way random effects model when the intraclass correlation is negative. Thus covariance parameters have an advantage over variance components in that the former are more generally applicable, as was pointed out in Dawid (1988). We will exploit this advantage in the formulation provided below.

Table 1. Covariance structure.

| observations | Covariance | |
|---|---|---|
| $(i,j)$ and $(i,j)$ | $\sigma^2$ | $2\sigma_g^2+\sigma_s^2+2\sigma_d^2+\sigma_r^2$ |
| $(i,j)$ and $(l,m)$ | $\sigma^2\rho_0$ | 0 |
| $(i,j)$ and $(j,i)$ | $\sigma^2\rho_1$ | $2\sigma_g^2+\sigma_s^2-2\sigma_d^2-\sigma_r^2$ |
| $(j,i)$ and $(i,l)$ | $\sigma^2\rho_2$ | $\sigma_g^2-\sigma_d^2$ |
| $(i,j)$ and $(i,l)$ | $\sigma^2\rho_3$ | $\sigma_g^2+2\sigma_{gd}+\sigma_d^2$ |
| $(i,j)$ and $(l,j)$ | $\sigma^2\rho_4$ | $\sigma_g^2-2\sigma_{gd}+\sigma_d^2$ |

## 3. Relation to Models in Genetics and Psychology

Model (1), along with some close variants, has been proposed in genetics, social psychology and psychophysics, with very little cross-referencing among the fields. Model (b) in Cockerham and Weir (1977) is identical to (1). The Social Relations Model (SRM) presented in Kenny (1994, p.232) (see also the references therein), which is the Model (a) of Cockerham and Weir (1977) under different notation, is a linear transformation of (1). Under the SRM, $y_{(i,j)}$ is decomposed into random effects $a$, $b$, and $c$, with

$$y_{(i,j)} = \mu + a_i + b_j + c_{ij}. \tag{2}$$

The terms in (1) and (2) are related by $a_i = g_i + d_i$, $b_j = g_j - d_j$, and $c_{ij} = s_{ij} + r_{ij}$. The parameters in (2) are the variances of $a$, $b$ and $c$, the covariance between $a_i$ and $b_i$, and the covariance between $c_{ij}$ and $c_{ji}$. Models (2) and (1) are equivalent in the sense that both generate the same pattern in, and impose the same constraints on, the covariance structure of the entire data vector. Furthermore, there is a one-to-one linear relationship between the parameters in (2) and (1).

A third formulation is the bio model given in Cockerham and Weir (1977), see also Lynch and Walsh (1998, p.605). The bio model differs slightly from (1) and (2), and has the form

$$y_{(i,j)} = \mu + n_i + n_j + t_{ij} + m_i + p_j + k_{ij}, \tag{3}$$

where $n_i$ and $n_j$ represent nuclear contributions, $m_i$ the extranuclear maternal effect, and $p_j$ the extranuclear paternal effect. The reader is referred to Cockerham and Weir (1977) for more details. Table 2 shows the relations among the parameters in Models (1), (2) and (3). Essentially the same information is presented by Cockerham and Weir (1977, p.189).

Table 2. Parameters in Models (1), (2) and (3)

| Diallel | SRM | Bio |
|---|---|---|
| $\sigma_g^2 - \sigma_d^2$ | $\sigma_{ab}$ | $\sigma_n^2$ |
| $2\sigma_d^2 - 2\sigma_{gd}$ | $\sigma_b^2 - \sigma_{ab}$ | $\sigma_p^2$ |
| $2\sigma_d^2 + 2\sigma_{gd}$ | $\sigma_a^2 - \sigma_{ab}$ | $\sigma_m^2$ |
| $\sigma_s^2 - \sigma_r^2$ | $\sigma_{cc'} (= cov(c_{ij}, c_{ji}))$ | $\sigma_t^2$ |
| $2\sigma_r^2$ | $\sigma_c^2 - \sigma_{cc'}$ | $\sigma_k^2$ |

From Table 2 it is clear that Model (3) differs from Models (1) and (2) by putting additional constraints on the covariance parameters in Table 1. The three models have received independent attention because they represent distinct situations. It is true, for instance, that Models (1) and (2) are statistically equivalent, but they tend to suggest different mechanisms. In the diallel model, outcomes arise from a composition of effects that are symmetrical with respect to position ($g_i$ and $s_{ij}$), along with effects due to order of position ($d_i$ and $r_{ij}$). The model has been used extensively in genetics (Simms and Triplett (1996)), and in the field of psychophysics Bechtel (1967, 1971) derived two very similar models for analyzing ratings of all possible paired compositions and contrasts of stimuli.

Model (2), the SRM, has been advocated by Kenny (1994) to analyze social interactions. A common task in a social psychology experiment occurs when a perceiver rates a target on some dimension of interest. The rating is affected by the perceiver's disposition to give certain ratings, along with the target's projected level on the dimension. Here the diallel model is not the most natural as the perceiver and target effects are conceptually distinct: a personality rating, for example, is not a composition of two individual effects of the same kind. Lev and Kinder (1957) and Lashley and Bond (1997) describe other observational settings where the SRM model is a better choice than the diallel model.

Finally, the bio model has been popular in the genetics literature (Husband and Gurney (1998), Lipow and Wyatt (1999), Motten and Stone (2000)). Here the emphasis has been on the estimation of nuclear effects, and extranuclear effects attributable to maternal and paternal sources. Although the preference for particular forms of the variance component models and their associated parameterizations depends on the scientific context, we believe that one parameterization is clearly most convenient for likelihood-based inference for dyadic data. We call this parameterization canonical, and now formally introduce it.

## 4. Likelihood-Based Inference for Dyadic Data

Along with statistical models for dyadic data, methods of statistical inference have also been extensively documented. Both Cockerham and Weir (1977) and Warner, Kenny and Stoto (1979) contain quadratic point estimates and schemes of hypothesis testing for the models they postulate. Although their point estimates based on the first two moments can be justified without making any distributional assumptions, some of their proposed procedures of statistical inference, such as F-tests, can be established by assuming normality. So far no systematic theory of statistical inference based on normal likelihood, analogous to that for the ordinary analysis of variance, seems to be available to researchers in the relevant substantive areas and applied statisticians. In this section we offer such a systematic treatment.

First, we provide a simple expression for the likelihood, based on the covariance structure in Table 1 under the normal distribution, when all pairs of $N$ units are observed. This is called a diallel design (Cockerham and Weir (1977)) in the genetics literature and a round-robin design in the psychological literature. The likelihood for the situations in which there are repeated measures is also provided. From those expressions of likelihood, we next give the maximum likelihood estimates of the parameters, and the distribution of sufficient statistics which can be used to derive test statistics. Bayesian inference for parameters is then discussed. Finally, algorithms for maximum likelihood computation when there are missing data are outlined.

### 4.1. The likelihood function

Although not obvious from casual inspection, it is not difficult to verify directly that the covariance matrix of the vector of observations on all the $N(N-1)$ dyads that can be formed with $N$ units, with elements specified in Table 1, can be expressed as a linear combination of a set of known matrices such that its inverse is also a linear combination of the same set of matrices, with the coefficients being simple functions of those of the original matrix. Specifically,

let $\Sigma$ denote a covariance matrix having the pattern given in Table 1, then

$$\Sigma = \lambda_u E_u + \lambda_s E_s + \lambda_r E_r + \lambda_g E_g + \lambda_d E_d + \lambda_{gd}\Delta_{gd}, \tag{4}$$

$$\Sigma^{-1} = \lambda_u^{-1}E_u + \lambda_s^{-1}E_s + \lambda_r^{-1}E_r + (\lambda_g\lambda_d - \lambda_{gd}^2)^{-1}(\lambda_d E_g + \lambda_g E_d - \lambda_{gd}\Delta_{gd}). \tag{5}$$

A patterned covariance matrix described by Table 1 is said to be in its canonical form when itself and its inverse are expressed in the form of (4) and (5), and the parameters in those expressions are called canonical parameters. The $E$ matrices in (4) are orthogonal projections onto the subspaces arising from fitting (1) as a fixed effect model, with the correspondence made explicit through the choice of subscripts. The matrix $\Delta_{gd}$, as a linear operator, sends the range of $E_g$ onto that of $E_d$, the range of $E_d$ onto that of $E_g$ and all vectors in the orthogonal complement of the sum of $E_d$ and $E_g$ to 0. The elements of the matrices in (4) are displayed in Table 3.

Table 3. The elements of matrices.

| index | $\Sigma$ | $E_u$ | $E_g$ | $E_s$ | $E_d$ | $E_r$ | $\Delta_{gd}$ |
|---|---|---|---|---|---|---|---|
| $(i,j),(i,j)$ | $\sigma^2$ | $\frac{1}{N(N-1)}$ | $\frac{1}{N}$ | $\frac{1}{2}-\frac{1}{N}-\frac{1}{N(N-1)}$ | $\frac{1}{N}$ | $\frac{1}{2}-\frac{1}{N}$ | 0 |
| $(j,i),(i,j)$ | $\sigma^2\rho_1$ | $\frac{1}{N(N-1)}$ | $\frac{1}{N}$ | $\frac{1}{2}-\frac{1}{N}-\frac{1}{N(N-1)}$ | $-\frac{1}{N}$ | $-\frac{1}{2}+\frac{1}{N}$ | 0 |
| $(i,j),(j,k)$ | $\sigma^2\rho_2$ | $\frac{1}{N(N-1)}$ | $\frac{N-4}{2N(N-2)}$ | $-\frac{N-4}{2N(N-2)}-\frac{1}{N(N-1)}$ | $-\frac{1}{2N}$ | $\frac{1}{2N}$ | 0 |
| $(i,j),(i,k)$ | $\sigma^2\rho_3$ | $\frac{1}{N(N-1)}$ | $\frac{N-4}{2N(N-2)}$ | $-\frac{N-4}{2N(N-2)}-\frac{1}{N(N-1)}$ | $\frac{1}{2N}$ | $-\frac{1}{2N}$ | $\frac{1}{\sqrt{N(N-2)}}$ |
| $(i,j),(k,j)$ | $\sigma^2\rho_4$ | $\frac{1}{N(N-1)}$ | $\frac{N-4}{2N(N-2)}$ | $-\frac{N-4}{2N(N-2)}-\frac{1}{N(N-1)}$ | $\frac{1}{2N}$ | $-\frac{1}{2N}$ | $-\frac{1}{\sqrt{N(N-2)}}$ |
| $(i,j),(k,l)$ | $\sigma^2\rho_0$ | $\frac{1}{N(N-1)}$ | $-\frac{2}{N(N-2)}$ | $\frac{2}{N(N-2)}-\frac{1}{N(N-1)}$ | 0 | 0 | 0 |

From the simple expression (5) for the inverse, and Table 3 for the matrix elements, we can write down the complete data loglikelihood (integrating out $\mu$) in terms of canonical parameters as

$$\begin{aligned} &-\frac{(N-1)(N-2)}{4}\ln\lambda_r - \frac{N(N-3)}{4}\ln\lambda_s - \frac{N-1}{2}\ln[\lambda_g\lambda_d - \lambda_{gd}^2] \\ &-\frac{tr(E_r\mathbf{y}\mathbf{y}')}{2\lambda_r} - \frac{tr(E_s\mathbf{y}\mathbf{y}')}{2\lambda_s} - \frac{\lambda_d tr(E_g\mathbf{y}\mathbf{y}') + \lambda_g tr(E_d\mathbf{y}\mathbf{y}') - \lambda_{gd}tr(\Delta_{gd}\mathbf{y}\mathbf{y}')}{2[\lambda_g\lambda_d - \lambda_{gd}^2]}. \end{aligned} \tag{6}$$

Thus, the MLE's of the canonical parameters are:

$$\begin{aligned} &\hat\lambda_r = \frac{2tr(E_r\mathbf{y}\mathbf{y}')}{(N-1)(N-2)}, \quad \hat\lambda_s = \frac{2tr(E_r\mathbf{y}\mathbf{y}')}{N(N-3)}, \quad \hat\lambda_g = \frac{tr(E_g\mathbf{y}\mathbf{y}')}{N-1}, \\ &\hat\lambda_d = \frac{tr(E_d\mathbf{y}\mathbf{y}')}{N-1}, \qquad \hat\lambda_{gd} = \frac{tr(\Delta_{gd}\mathbf{y}\mathbf{y}')}{2(N-1)}. \end{aligned} \tag{7}$$

From the above MLE's for the canonical parameters we can obtain MLE's for the covariance parameters by using Table 3, which expresses covariance parameters as linear combinations of canonical parameters whose coefficients can be found in the corresponding columns. The variance parameters in Models (1), (2) and (3) may then be obtained by using Tables 1 and 2. However, although the MLE's for the canonical parameters will automatically satisfy the necessary and sufficient conditions for the covariance matrix $\Sigma$ to be nonnegative definite, which are the nonnegativity of $\lambda_s$, $\lambda_r$, and the nonnegative definiteness of $\Lambda_{gd}$ (see (17)), there is no guarantee that using this approach the estimated variance components will have nonnegative variances and nonnegative definite covariance matrices in Models (1), (2) and (3). Various additional constraints need to be imposed on the canonical parameters. In the specific case of Model (3), such constraints can be found in Section 5.

### 4.2. The likelihood function with repeated measures

Warner, Kenny and Stoto (1979) considered round-robin designs with repeated measures on each dyad. Their model amounts to adding an error term to Model (2), or equivalently (in the statistical sense) Model (1), that is correlated between reciprocal dyads, and thus can be expressed as follows:

$$y_{(i,j)k} = \mu + g_i + g_j + s_{ij} + d_i - d_j + r_{ij} + e_{ijk}, \tag{8}$$

where $(i,j)k$ denotes the $k$th repeated measure on dyad $(i,j)$, $var(e_{ijk}) = \sigma_e^2$ and $cov(e_{ijk}, e_{jik}) = \sigma_e^2\rho$. Expressions (4) and (5) can also be extended for the covariance matrix $\Sigma$ (and its inverse) for the balanced round-robin design with $K$ observations on all the $N(N-1)$ dyads that can be formed with $N$ units under Model (8). Specifically, the $KN(N-1) \times KN(N-1)$ matrix $\Sigma$ and its inverse can be expressed as

$$\Sigma = \lambda_u E_u + \lambda_s E_s + \lambda_r E_r + \lambda_g E_g + \lambda_d E_d + \lambda_{gd}\Delta + \lambda_v E_v + \lambda_w E_w, \tag{9}$$

$$\begin{aligned}\Sigma^{-1} = {} & \lambda_u^{-1} E_u + \lambda_s^{-1} E_s + \lambda_r^{-1} E_r + (\lambda_g\lambda_d - \lambda_{gd}^2)^{-1}(\lambda_d E_g + \lambda_g E_d - \lambda_{gd}\Delta_{gd}) \\ & + \lambda_v^{-1} E_v + \lambda_w^{-1} E_w,\end{aligned} \tag{10}$$

where the elements of $E_v$ and $E_w$, both orthogonal projections, are tabulated in Table 4, and the rest of the matrices are defined as in Table 3 with each entry replaced by a $K \times K$ matrix in which all the elements are $\frac{1}{K}\times$(the original entry). In terms of group invariance, covariance matrices of the form (9) are precisely those invariant under the permutations on the sampling units and pairs of repeated measurement on reciprocal dyads.

Table 4. The elements of $E_v$ and $E_w$.

| $(i,j)k$, | $(i,j)k$ | $(j,i)k$ | $(i,j)k'$ | $(j,i)k'$ | other |
|---|---|---|---|---|---|
| $E_v$ | $\frac{1}{2}-\frac{1}{2K}$ | $\frac{1}{2}-\frac{1}{2K}$ | $-\frac{1}{2K}$ | $-\frac{1}{2K}$ | 0 |
| $E_w$ | $\frac{1}{2}-\frac{1}{2K}$ | $-\frac{1}{2}+\frac{1}{2K}$ | $-\frac{1}{2K}$ | $\frac{1}{2K}$ | 0 |

Table 5 gives the relation between the canonical parameters and the variance components for Model (8), from which the relation between the parameters in (1) and (5) can be obtained by setting $K = 1$ and $\sigma_e^2 = 0$. The complete data loglikelihood in terms of canonical parameters is

$$\begin{aligned} &-\frac{(N-1)(N-2)}{4}\ln\lambda_r - \frac{N(N-3)}{4}\ln\lambda_s - \frac{N-1}{2}\ln[\lambda_g\lambda_d - \lambda_{gd}^2] \\ &-\frac{N(N-1)(K-1)}{4}(\ln\lambda_w + \ln\lambda_v) - \frac{tr(E_v\mathbf{y}\mathbf{y}')}{2\lambda_v} - \frac{tr(E_w\mathbf{y}\mathbf{y}')}{2\lambda_w} - \frac{tr(E_r\mathbf{y}\mathbf{y}')}{2\lambda_r} \\ &-\frac{tr(E_s\mathbf{y}\mathbf{y}')}{2\lambda_s} - \frac{\lambda_d tr(E_g\mathbf{y}\mathbf{y}') + \lambda_g tr(E_d\mathbf{y}\mathbf{y}') - \lambda_{gd} tr(\Delta_{gd}\mathbf{y}\mathbf{y}')}{2[\lambda_g\lambda_d - \lambda_{gd}^2]} \end{aligned} \tag{11}$$

From the above loglikelihood, the MLE's are:

$$\hat{\lambda}_v = \frac{2tr(E_v\mathbf{y}\mathbf{y}')}{N(N-1)(K-1)} \quad \hat{\lambda}_w = \frac{2tr(E_w\mathbf{y}\mathbf{y}')}{N(N-1)(K-1)} \quad \hat{\lambda}_{gd} = \frac{tr(\Delta_{gd}\mathbf{y}\mathbf{y}')}{2(N-1)}, \tag{12}$$

with the rest having the same formal expressions as in (7). The MLE's automatically satisfy the necessary and sufficient conditions on the canonical parameters for the entire covariance matrix to be nonnegative definite: that $\lambda_v$, $\lambda_w$, $\lambda_r$, $\lambda_s$ be nonnegative and $\Lambda_{gd}$ (see (17)) be nonnegative definite.

Table 5. Repeated measures parameters.

| Canonical | Variance |
|---|---|
| $\lambda_g$ | $2K(N-2)\sigma_g^2 + \lambda_s$ |
| $\lambda_s$ | $2K\sigma_s^2 + (1+\rho)\sigma_e^2$ |
| $\lambda_d$ | $2KN\sigma_d^2 + \lambda_r$ |
| $\lambda_r$ | $2K\sigma_r^2 + (1-\rho)\sigma_e^2$ |
| $\lambda_v$ | $(1+\rho)\sigma_e^2$ |
| $\lambda_w$ | $(1-\rho)\sigma_e^2$ |
| $\lambda_{gd}$ | $2K\sqrt{N(N-2)}\sigma_{gd}$ |

### 4.3. The complete data inference

Statistical inference beyond point estimation for variance components under round-robin (diallel) designs can be based on the joint distribution of the sufficient

statistics in (11) (or (6) when $K = 1$). In what follows we work with the more general (11). The distribution of the sufficient statistics, which are quadratic forms of the data vector, are given below.

$$\begin{aligned}
SS_w &= \mathbf{y}'E_w\mathbf{y} = tr(E_w\mathbf{y}\mathbf{y}') \sim \lambda_w\chi^2_{\frac{1}{2}(K-1)N(N-1)},\\
SS_v &= \mathbf{y}'E_v\mathbf{y} = tr(E_v\mathbf{y}\mathbf{y}') \sim \lambda_v\chi^2_{\frac{1}{2}(K-1)N(N-1)},\\
SS_g &= \mathbf{y}'E_g\mathbf{y} = tr(E_g\mathbf{y}\mathbf{y}') \sim \lambda_g\chi^2_{(N-1)},\\
SS_s &= \mathbf{y}'E_s\mathbf{y} = tr(E_s\mathbf{y}\mathbf{y}') \sim \lambda_s\chi^2_{\frac{1}{2}N(N-3)},\\
SS_d &= \mathbf{y}'E_d\mathbf{y} = tr(E_d\mathbf{y}\mathbf{y}') \sim \lambda_d\chi^2_{(N-1)},\\
SS_r &= \mathbf{y}'E_r\mathbf{y} = tr(E_r\mathbf{y}\mathbf{y}') \sim \lambda_r\chi^2_{\frac{1}{2}(N-1)(N-2)},\\
\begin{pmatrix} SS_g & SC_{gd} \\ SC_{gd} & SS_d \end{pmatrix} &\sim W_2\left(\begin{pmatrix} \lambda_g & \lambda_{gd} \\ \lambda_{gd} & \lambda_d \end{pmatrix}, N-1\right),
\end{aligned} \tag{13}$$

where $SC_{gd} = \frac{1}{2}\mathbf{y}'\Delta_{gd}\mathbf{y} = \frac{1}{2}tr(\Delta_{gd}\mathbf{y}\mathbf{y}')$. We use the notation '$SS$' ('$SC$') for the quadratic forms because they have the same algebraic and distributional properties as sums of squares in analysis of variance or sums of cross products in multivariate analysis. The sums of squares add up to the total sum of squares just as in an ordinary analysis of variance. Since the above distributional results can be read off the loglikelihood (11), formal derivations are omitted. Based on the distributions of the sufficient statistics, appropriate procedures can be derived for statistical inference with regard to the variance components. Wong (1982) also addressed the likelihood-based inference for Model (8), but did not obtain closed form MLE's (7) (or (12)) and simple expressions of loglikelihoods (11) (or (6)). Nor did he obtain all the sufficient statistics and their distributions. Cockerham and Weir (1977), working with the special case of $\rho = 0$, correctly pointed out that some of the sums of squares in their paper are correlated. However, their method was not based on the likelihood function, and therefore they did not give the joint distribution of the sums of squares.

When $\rho = 0$, as is the case in most diallel designs in genetics, $\lambda_v = \lambda_w$, and $SS_w$ and $SS_v$ can be pooled to become the 'within dyad' sum of squares. Let $\lambda_e = \lambda_v = \lambda_w$, $SS_e = SS_w + SS_v$, and $E_e = E_w + E_v$. The loglikelihood for this special case is

$$\begin{aligned}
&-\frac{(N-1)(N-2)}{4}\ln\lambda_r - \frac{N(N-3)}{4}\ln\lambda_s - \frac{N-1}{2}\ln[\lambda_g\lambda_d - \lambda_{gd}^2]\\
&-\frac{N(N-1)(K-1)}{2}\ln\lambda_e - \frac{tr(E_e\mathbf{y}\mathbf{y}')}{2\lambda_e} - \frac{tr(E_r\mathbf{y}\mathbf{y}')}{2\lambda_r} - \frac{tr(E_s\mathbf{y}\mathbf{y}')}{2\lambda_s}\\
&-\frac{\lambda_d tr(E_g\mathbf{y}\mathbf{y}') + \lambda_g tr(E_d\mathbf{y}\mathbf{y}') - \lambda_{gd}tr(\Delta_{gd}\mathbf{y}\mathbf{y}')}{2[\lambda_g\lambda_d - \lambda_{gd}^2]}
\end{aligned} \tag{14}$$

From the above loglikelihood we can see that the sufficient statistics $SS_w = \mathbf{y}'E_w\mathbf{y}$ and $SS_v = \mathbf{y}'E_v\mathbf{y}$ collapse into $SS_e = SS_w + SS_v = \mathbf{y}'E_e\mathbf{y} \sim \lambda_e\chi^2_{(K-1)N(N-1)}$. The covariance matrices under the condition $\rho = 0$ are precisely those invariant under all permutations on the sampling units and on repeated measures independently within each dyad.

In the absence of $\rho$ the mean square associated with $SS_e$ is the 'error term', the maximum likelihood and unbiased estimator of $\sigma_e^2$ in Model (8). However, depending on the actual design, a number of degrees of freedom in $SS_e$ may need to be allocated to certain design effects, with the remainder serving as error term. For the diallel experiment example in Cockerham and Weir (1977), one degree of freedom is allocated to a 'block effect'.

### 4.4. The likelihood computation when there are missing data and covariates

Additional problems in the estimation of the parameters in Model (8) (and its special cases and variants) resulting from missing data and covariates can be handled by the EM algorithm (Dempster, Laird and Rubin (1977)) and its recent extensions (e.g., Meng and Rubin (1993), Liu and Rubin (1994)). Here the phrase 'missing data' is used in its most general sense, which includes the cases when certain observations are omitted by design, as well as the cases when data are lost by accident. Indeed, if we think along the lines of Rubin and Szatrowski (1982), the relevance of the results in this section is not restricted to dyadic data. The results are useful for any patterned covariance that can be 'imbedded', in the sense of Rubin and Szatrowski (1982), in the covariance structures (4) or (9). In this section we only sketch the steps of the EM algorithm, without addressing the issue of efficient implementation. The reader is recommended to consult Liu and Rubin (1994), Liu, Rubin and Wu (1998), Meng and van Dyk (1997, 1998).

The EM algorithm for maximum likelihood estimation of the parameters based on the loglikelihood (11), when there are missing data, is sketched below. Specialization to Model (6) is straightforward. The t-th E step:

$$\mathbf{V}^{(t)} = \mathbf{y}^{*(t)}(\mathbf{y}^{*(t)})' + \begin{pmatrix} \hat{\boldsymbol{\Sigma}}^{(t)}_{mis,mis} - \hat{\boldsymbol{\Sigma}}^{(t)}_{mis,obs}(\hat{\boldsymbol{\Sigma}}^{(t)}_{obs,obs})^{-1}\hat{\boldsymbol{\Sigma}}^{(t)}_{obs,mis} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{15}$$

$$\begin{cases} \mathbf{y}^{*(t)}_{mis} = \hat{\mu}^{(t)}J_{mis} + \hat{\boldsymbol{\Sigma}}^{(t)}_{mis,obs}(\hat{\boldsymbol{\Sigma}}^{(t)}_{obs,obs})^{-1}(\mathbf{y}_{obs} - \hat{\mu}^{(t)}J_{obs}) \\ \mathbf{y}^{*(t)}_{obs} = \mathbf{y}_{obs} \qquad (J \ = \ \text{summing vector}) \end{cases}, \tag{16}$$

where $\mathbf{y}^{*(t)}_{mis}$ is the conditional expectation of $\mathbf{y}_{mis}$, and $\mathbf{V}^{(t)}$ is the conditional expectation of $\mathbf{y}\mathbf{y}'$ given $\mathbf{y}_{obs}$ at the values of parameters obtained from the previous M step. The (t+1)-st M step: use (12) and (7) with $\mathbf{y}\mathbf{y}'$ replaced by $\mathbf{V}^{(t)}$

to obtain the new estimates of the canonical parameters; the new estimate for the mean parameter is $\hat{\mu}^{(t+1)} = J'\mathbf{y}^{*(t)}$. For methods of evaluating the asymptotic variance-covariance matrix of the estimates obtained by EM algorithm, see Meng and Rubin (1991), van Dyk, Meng and Rubin (1995), and Oakes (1999). Note that, as in Little and Rubin (1987), 'mis' and 'obs' refer to the set of indices of missing and observed elements of data vector, respectively.

For special patterns of fixed effects superimposed on the covariance structure, closed form solutions could still be possible. When arbitrary covariates are incorporated into Model (12), MLE's can be obtained from the ECM algorithm (Meng and Rubin (1993)). The loglikelihood in the presence of fixed effects represented by model matrix $X$ and coefficients $\beta$ can be obtained from (11) by replacing $\mathbf{y}$ with $(\mathbf{y} - \mathbf{X}\beta)$. The ECM algorithm can be implemented through the following modification of EM algorithm: (1) replace $\hat{\mu}^{(t)}J$ in the E step of the EM algorithm (16) with $\mathbf{y} - \mathbf{X}\hat{\beta}^{(t)}$; (2) replace the $\mathbf{V}^{(t)}$ in the M step of the EM algorithm with $\mathbf{V}^{(t)} - 2\mathbf{y}^{*(t)}(\mathbf{X}\hat{\beta}^{(t)})' + (\mathbf{X}\hat{\beta}^{(t)})(\mathbf{X}\hat{\beta}^{(t)})'$; (3) $\hat{\beta}^{(t+1)} = (\mathbf{X}'(\hat{\Sigma}^{(t+1)})^{-1}\mathbf{X})^{-1}\mathbf{X}'(\hat{\Sigma}^{(t+1)})^{-1}\mathbf{y}^{*(t)}$. The EM algorithms described in this section appear to be simpler than those in Wong (1982), due to the availability of closed form MLE's and the simple expression of loglikelihood.

### 4.5. Toward Bayesian inference for dyadic data

The loglikelihood (11) has five additive components. Four of the components correspond to the density of the chi-square distribution with scale parameters $\lambda_w$, $\lambda_v$, $\lambda_r$, $\lambda_s$, respectively. The fifth component corresponds to a Wishart density with scale parameter

$$\Lambda_{gd} = \begin{pmatrix} \lambda_g & \lambda_{gd} \\ \lambda_{gd} & \lambda_d \end{pmatrix}. \tag{17}$$

If we assign priors independently for the parameters $\lambda_w$, $\lambda_v$, $\lambda_r$, $\lambda_s$, and $\Lambda_{gd}$ associated with each of the five components of the likelihood, which seems to be a reasonable thing to do, then those parameters would be independent a posteriori. The standard reference prior for $\lambda_w$, $\lambda_v$, $\lambda_r$, and $\lambda_s$ is Jeffreys's prior, under which the posterior distributions of those parameters are $\lambda_w|\mathbf{y} \sim SS_w\chi^{-2}_{\frac{1}{2}(K-1)N(N-1)}$; $\lambda_v|\mathbf{y} \sim SS_v\chi^{-2}_{\frac{1}{2}(K-1)N(N-1)}$; $\lambda_r|\mathbf{y} \sim SS_r\chi^{-2}_{\frac{1}{2}(N-1)(N-2)}$; $\lambda_s|\mathbf{y} \sim SS_s\chi^{-2}_{\frac{1}{2}N(N-3)}$. There seems to be less agreement on the choice of reference prior for the scale parameters of a Wishart density, a topic that deserves further study. The reader is advised to consult recent references, such as Yang and Berger (1994), on this topic. In a subsequent numerical example, we choose for $\Lambda_{gd}$ the reference prior documented in Box and Tiao (1973, p.426).

The Bayesian approach can overcome difficulties with the frequentist approach with regard to the inference for some of the parameters. Examples of

such parameters are the variances $\sigma_a^2$ and $\sigma_b^2$, in the Social Relations Model, and the parameter $\sigma_n^2$ $(= \sigma_g^2 - \sigma_d^2)$ in Cockerham and Weir (1977). The inference for those parameters, especially beyond point estimation, is a hard problem within the frequentist framework, but can readily be dealt with from the Bayesian perspective. The posterior distribution of those parameters can easily be found, through simple simulation if necessary, given the joint distribution of the canonical parameters. The constraints induced by the nonnegativity of variance components are also easier to cope with in the Bayesian framework (Box and Tiao (1973, p.67)).

## 5. Some Practical Implications of the Theoretical Advances Illustrated by Numerical Examples

The likelihood-based and Bayesian methods of statistical inference for dyadic data developed in this paper is not only of theoretical interest, but also of practical significance. In this section we illustrate the application of the methods proposed in the previous sections. The data set used for the illustration is displayed in Appendix C of Cockerham and Weir (1977), which contains the flowering times in days of crosses from eight inbred lines, originally from Hayman (1954). We perform two statistical procedures that are beyond what can be provided by the methods developed in Cockerham and Weir (1977).

The first procedure is a test of the null hypothesis $\sigma_m^2 = \sigma_p^2$ in the bio model, which is exact under the assumption of normality. From Tables 2 and 5, this hypothesis is equivalent to $\lambda_{gd} = 0$. Based on the last equation in (13), an obvious test statistic is $h = SC_{gd}/\sqrt{SS_g SS_d}$, which is distributed as an ordinary sample correlation coefficient with its corresponding population value equal to $\lambda_{gd}/\sqrt{\lambda_g \lambda_d}$. In particular the null distribution of $\sqrt{N-1}h/\sqrt{1-h^2}$ is a t-distribution with $N-1$ degrees of freedom. Cockerham and Weir (1977) provided a test statistic for the null hypothesis $\sigma_m^2 = \sigma_p^2$ for the factorial data structure, but not the diallel data structure, due perhaps to the limitations of the methods of symmetrical products (Koch (1967)) and of squares of symmetrical differences (Koch (1968)) employed therein. For numerical illustration, we use data in Appendix C of Cockerham and Weir (1977), from which we obtain $SC_{gd} = -67.635$, $SS_g = 1225.312$, and $SS_d = 374.026$. Therefore, $h = -0.1$ and the corresponding t statistic is equal to $-0.266$, which indicates that $\lambda_{gd}$, and hence $\sigma_m^2 - \sigma_p^2$, is not significantly different from 0.

The second procedure is an empirical test for the plausibility of the bio model itself. For any covariance matrix whose pattern is specified in (9), or the special case of $\lambda_v = \lambda_w$ as a result of the appropriate group invariance, the only constraints on the canonical parameters imposed by the nonnegative definiteness are that they are all nonnegative and $\Lambda_{gd}$ is nonnegative definite. Any variance

component model would impose additional constraints. Tables 2 and 5 tell us that the Cockerham-Weir model (3) requires more stringent constraints than the other two variance component models discussed in this paper. It would therefore be natural to ask how likely do the constraints imposed by (3), or its repeated measures version, hold given a set of data. A very simple way to address this question is to calculate the posterior probability that the constraints are satisfied. If we choose the prior as described in Section 4.5, and use the reference prior in Box and Tiao (1973, Chapter 8) for $\Lambda_{gd}$, then $\Lambda_{gd}^{-1}|\mathbf{y} \sim W(S_{gd}^{-1}, N-1)$, and the distributions of the other parameters are given in Section 4.5. Now it is straightforward to calculate, via simulation, the probability that the following constraints imposed by the bio model are met: $(\lambda_g - \lambda_s)/2K(N-2) - (\lambda_d - \lambda_r)/2KN \geq 0$, $-(\lambda_d - \lambda_r)/N \leq \lambda_{gd}/\sqrt{N(N-2)} \leq (\lambda_d - \lambda_r)/N$, $\lambda_s \geq \lambda_r$, $\lambda_r \geq \lambda_e$. Ten thousand samples were obtained from the posterior distribution using Splus on a Sun workstation. Out of those 10000 samples 5334 meet the constraints, hence the posterior odds is 1.143 : 1 in favor of the bio model.

The two examples in this section cover only a small fraction of the range of potential applications that can be of utility to substantive fields. In particular, the scheme for Bayesian inference is general enough to address statistical inference for all the parameters, including those of scientific interest in the Social Relations Model, the diallel model, and the bio model. There is clearly a need for inferential tools for a variety of parameters, see for example Husband and Gurney (1998, p.175) in the case of bio model. Sometimes it may be desirable to adapt the general method for specific purposes, and this is a good topic for future investigation.

## 6. Discussion

The subject of dyadic data structure has a long history in statistics. Frank Yates addressed it in his article in the first volume of Heredity, in which (1) appeared as a fixed effects model. Since then the problem has emerged in the statistical literatures on judgement and perception, sports, genetics, and social relations. Although there have been obvious similarities among the various formulations, they have been largely independently developed and ad hoc due to the lack of a unified framework for modelling and inference. Methods of statistical inference grounded on a sound theoretical foundation and having the capability for handling general patterns of missing data and unbalance, which commonly occur even in the best designed experiments, are lacking (Cockerham and Weir (1977), Husband and Gurney (1998)). Our work bridges these gaps by showing the underlying connections among the various approaches. The likelihood-based inference numerically reproduces the point estimation obtained in Warner, Kenny

and Stoto (1979), Lev and Kinder (1957), Bechtel (1967, 1971), and Cockerham and Weir (1977). However, we go beyond point estimation by introducing streamlined statistical procedures for joint inference for all the parameters based on likelihood and Bayesian methodology, and with full missing data capabilities. Generally, we hope that the availability of the powerful and convenient statistical methods will encourage a wider application of dyadic data structure in designing empirical studies and in formulating scientific theories. We also hope that the technical aspect of our work will be a useful experience for the study of statistical inference for other data structures under more general exchangeability and distributional conditions. In fact, the results in Andersson (1975), which are succinctly summarized in Perlman (1987), indicate that decompositions in the form of (4), which serves as the starting point of our work, are available for all covariance structures invariant under some group actions.

## Acknowledgements

The first author thanks the NSF for a travel grant that enabled him to present part of this paper at CLAPEM'98, and the NIH for support through grants M01 RR00044 and P01 DE13539. The second author thanks Educational Testing Service for support from a Harold Gulliksen fellowship. Both authors would like to thank Professor Donald B. Rubin for his generous help.

## References

Andersson, S. A. (1975). Invariant normal models. *Ann. Statist.* **3**, 132-154.

Bond, C. F. and Lashley, B. R. (1996). Round-robin analysis of social interactions: Exact and estimated standard errors. *Psychometrika* **61**, 303-311.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* Addison-Wesley, Reading, MA.

Bechtel, G. G. (1967) The analysis of variance and pairwise scaling. *Psychometrika* **32**, 47-65.

Bechtel, G. G. (1971) A dual scaling analysis for paired compositions. *Psychometrika* **36**, 135-154.

Cockerham, C. C. and Weir, B. S. (1977). Quadratic analyses of reciprocal crosses. *Biometrics* **33**, 187-204.

Dawid, A. P. (1988). Symmetry models and hypotheses for structured data layouts (with discussion). *J. R. Statist. Soc. B* **50**, 1-34.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Hayman, B. I. (1954). The theory and analysis of diallel crosses. *Genetics* **39**, 789-809.

Husband, B. C. and Gurney, J. E. (1998). Offspring fitness and parental effects as a function of inbreeding in Epilobium angustifolium (Onagraceae). *Heredity* **80**, 173-179.

Iacobucci, D. and Wasserman, S. (1987). Dyadic social interactions. *Psych. Bull.* **102**, 293-306.

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* Wiley, New York.

Kenny, D. A. (1994). *Interpersonal Perception.* Guilford Publications, Inc., New York.

Koch, G. G. (1967). A general approach to the estimation of variance components. *Technometrics* **9**, 93-118.

Koch, G. G. (1968). A general approach to the estimation of variance components. *Technometrics* **10**, 551-558.

Kraemer, H. C. and Jacklin, D. N. (1979). Statistical analysis of dyadic social behavior. *Psych. Bull.* **86**, 217-224.

Lashley, B. R. and Bond, C. F. (1997). Significance tests for round-robin data. *Psychological Methods* **2**, 278-291.

Lev, J. and Kinder, E. (1957). New analysis of variance formulas for treating data from mutually paired subjects. *Psychometrika* **22**, 1-15.

Lipow, S. R. and Wyatt, R. (1999). Diallel crosses reveal patterns of variation in fruit-set, seed mass, and seed number in Asclepias incarnata. *Heredity* **83**, 310-318.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

Liu, C. and Rubin, D. B. (1994). The ECME algorithm: a simple extension of EM and ECM with facter monotone convergence. *Biometrika* **81**, 633-648.

Liu, C., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* **85**, 755-770.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc., Sunderland, MA.

Meng, X-L. and Rubin, D. B. (1991). Using the EM algorithm to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.* **86**, 899-909.

Meng, X-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80**, 267-278.

Meng, X-L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meng, X-L. and van Dyk, D. A. (1998). Fast EM implementations for mixed-effects models. *J. Roy. Statist. Soc. Ser. B* **60**, 559-578.

Motten, A. F. and Stone, J. L. (2000). Heritability of stigma position and the effect of stigma-anther separation of outcrossing in a predominantly self-fertilizing weed, Datura Stramonium (Solanaceae). *Am. J. Bot.* **87**, 339-347.

Oakes, D. (1999). Direct calculation of the information matrix via the EM algorithm *J. Roy. Statist. Soc. Ser. B* **61**, 479-482.

Perlman, M. D. (1987). Comment: Group symmetry covariance models. *Statist. Sci.* **2**, 421-425.

Rubin, D. B. and Szatrowski, T. H. (1982). Finding maximum likelihood estimates of patterned covariance matrices by the EM algorithm. *Biometrika* **69**, 657-660.

Simms, E. L. and Triplett, J. K. (1996). Paternal effects in inheritance of a pathogen resistance trait in Ipomoea Purpurea. *Evolution* **50**, 2178-2186.

van Dyk, D. A., Meng X-L. and Rubin, D. B. (1995). Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statist. Sinica* **5**, 55-75.

Warner, R. M., Kenny, D. A. and Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *J. Person. Soc. Psych.* **37**, 1742-1757.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis.* Cambridge University Press.

Wong, G. (1982). Round robin analysis of variance via maximum likelihood. *J. Amer. Statist. Assoc.* **77**, 714-724.

Wright, A. J. (1985). Diallel designs, analyses, and reference populations. *Heredity* **54**, 307-311.

Xu, Z. C. and Zhu, J. (1999). An approach for predicting heterosis based on an additive, dominance and additive x additive model with environment interaction. *Heredity* **82**, 510-517.

Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22**, 1195-1211.

Yates, F. (1947). Analysis of data from all possible reciprocal crosses between a set of parental lines. *Heredity* **1**, 287-301.

Zhu, J. and Weir, B. S. (1996a). Diallel analysis for sex-linked and maternal effects. *Theor. Appl. Genet.* **92**, 1-9.

Zhu, J. and Weir, B. S. (1996b). Mixed approaches for diallel analysis based on a bio-model. *Genet. Res. Camb.* **68**, 233-240.

Department of Biostatistics, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, U.S.A.

E-mail: liheng@bst.rochester.edu

College of Health and Human Development, The Pennsylvania State University, University Park, PA 16802-6501, U.S.A.

E-mail: loken@psu.edu

(Received June 2000; accepted October 2001)