Model formulae for homologous factors
by

Peter McCullagh
Department of Statistics
University of Chicago

February 1998

1. Introduction

Two factors A and B are said to be homologous, or isomorphic, if they have the same set of levels. That is to say, not only do A and B have the same number of levels, but the sets of labels are identical, or in a natural 1–1 correspondence. The list of examples in the following section serves to illustrate three facts. First, homologous factors are fairly common in applications. Second, the usual sorts of factorial models are insufficient to describe the types of relationships that typically occur when homologous factors are present. Third, the familiar and widely accepted marginality condition, that interactions cannot be present without main effects, must be modified for homologous factors.

The standard set of symbolic operators described by Wilkinson and Rogers (1972), and widely implemented in computer packages, is inadequate for homologous factors. The purpose of this paper is to describe a minimal extension capable of describing the most common non-factorial models that occur in designs that have homologous factors. The extension is in fact algebraically complete only for two homologous factors, but it seems to include the most interesting spaces for up to four such factors.

2. Examples

2.1 *Citation analysis*

A common practice in citation analyses is to begin by tabulating the number of citations y_{ij} in journal i to journal j over a given period (Stigler, 1994). In the present framework, A is the factor 'citing journal' and B is the homologous factor 'cited journal.' Within a particular subject area, two phenomena are usually apparent, a strong tendency for self-citations, and an academic hierarchy reminiscent of social classes. The applied journals speak only to the methodological journals, the methodological journals speak only to the theoretical journals, and the theoretical journals speak only to god. A similar phenomenon occurs in geographical studies of migration, in which y_{ij} is the number of individuals, families, or other units, moving from region i to region j in a specified period. Short migrations within regions are much more numerous than longer migrations between regions. At the same time, certain regions become more popular, or economically more active, so that the relative migration rate from region i to region j depends on the difference in the growth rates of the two regions.

The null model in both cases is that the expected value of Y_{ij} is the same as the expected value of Y_{ji}. In other words, the table of expected values is symmetric. The non-null model, usually assumed to have the Bradley-Terry form,

$$\log \mu_{ij} = \phi_{ij} + \alpha_i - \alpha_j,$$

where $\phi_{ij} = \phi_{ji}$, is designed to yield a linear ranking of the journals or regions. Stigler refers to the asymmetry term $\alpha_i - \alpha_j$ as the imbalance of trade in intellectual influence between journals i and j. The matrix $\log(\mu_{ij}/\mu_{ji})$ is skew-symmetric by construction and additive by assumption:

$$\log \mu_{ij} - \log \mu_{ji} = 2\alpha_i - 2\alpha_j,$$

so the numbers α_i determine the linear ranking of journals or regions. A matrix μ satisfying this condition is said to be quasi-symmetric (Caussinus, 1965; McCullagh, 1982). In citation studies, quasi-symmetry is known as the 'export scores model' (Stigler, 1994); in migration studies, the term 'gravity model' is used. (Stewart 1948, Upton 1985).

2.2 *Plant breeding*

In plant breeding studies, males of several varieties are crossed with females of the same varieties. The response in such an experiment is a function on the varieties crossed with themselves, i.e. an element in the tensor product space $A^{\otimes 2}$, or $A \otimes B$, where the levels of A and B are the varieties. The Mendelian inheritance model implies that the offspring of male variety i crossed with female variety j is, in a statistical sense, genetically identical to male variety j crossed with female variety i. In other words, the expected response $\mu_{ij} = E(Y_{ij})$ is symmetric, a property here denoted by $\mu \in \mathrm{sym}^2(A, B)$. A stronger form of the model implies that the genetic contribution of a male is independent of the variety of the female. In practice, the stronger condition is interpreted as

$$g(\mu_{ij}) = \alpha_i + \alpha_j,$$

additive for some monotone transformation g, but with equal contributions from the two parents. This form of additive model is used implicitly by Yates (1947), and explicitly by Griffing (1956) after whom it is sometimes named. It is denoted here by $\mathrm{symadd}(A, B)$, or by $\mathrm{sym}(A + B)$ for brevity.

It should be emphasized that Mendel's laws do not imply additivity, and that favourable or unfavourable conjunctions of genes guarantee non-additivity in most cases. Nevertheless, this symmetric additive model often accounts for a large fraction of the systematic variation provided that incompatible pairs are excluded.

2.3 *Tournament analysis*

Consider a tournament in which teams or players compete with each other in pairs. The results of such a tournament can be compiled in a table indexed by the set of teams crossed with itself, with entries

$$y_{ij} = \begin{cases} 1 & \text{player } i \text{ beat player } j; \\ -1 & \text{player } j \text{ beat player } i; \\ 0 & \text{draw.} \end{cases}$$

By construction, y is skew-symmetric, so any model for y must focus on the skew-symmetric subspaces of functions on the square. Two slightly different models for the probabilities $\pi_{ij} = \mathrm{pr}(Y_{ij} = 1)$ are as follows:

$$\begin{aligned} \log(\pi_{ij}/\pi_{ji}) &= \alpha_i - \alpha_j \\ \mathrm{logit}\, \pi_{ij} &= \alpha'_0 + \alpha'_i - \alpha'_j \end{aligned}$$

The subspaces indicated by the expressions on the right are here denoted by $\mathrm{altadd}(A, B)$ and $1+\mathrm{altadd}(A, B)$ respectively. Note also that 1 is not a subspace of $\mathrm{altadd}(A, B)$, so these two expressions represent distinct nested subspaces. Because A and B are homologous, the absence of the intercept should not be regarded as a violation of Nelder's (1977) marginality criteria (McCullagh, 1998).

If draws do not occur, $\alpha'_0 = 0$ and the models coincide. Otherwise the models are distinct.

2.4 *Matched pairs design*

Consider a matched pairs design in which the response for each subject is a value in the finite set Ω, the set of response categories. In a study of identical twins reared apart, for example, the response might be some form of preference, for say, clothing, colour, music, political party or sexual orientation. In a medical context, the response might be a measure of pain relief or improvement in medical condition. In each of these cases, the response for each pair is a point in $\Omega \times \Omega$, and y_{rs}, the number of pairs responding (r, s), are the components of a vector in $A^{\otimes 2}$.

Absence of a treatment effect implies symmetry, i.e. $\mu \in \mathrm{sym}^2(A, B)$. Various non-null models have been derived or proposed including quasi-symmetry (Caussinus 1965; McCullagh 1982) and marginal homogeneity (Agresti, 1983). Marginal homogeneity is that subspace of $A^{\otimes 2}$ for which the expected number of treated members responding in category r is the same as the expected number of controls in category r, $\mu_{r\bullet} = \mu_{\bullet r}$. Marginal homogeneity and quasi-symmetry are invariant under permutation of the labels in Ω, that is to say group-invariant subspaces. McCullagh (1997) argues that models specified by subspaces ought to be monoid-invariant, a semi-group condition stronger than group-invariance. Quasi-symmetry passes this test, but marginal homogeneity does not. All of the models described in this paper, including all of the factorial models of Wilkinson and Rogers (1972), are monoid-invariant.

2.5 *Diallel cross design*

A slightly more complicated version of the plant-breeding example occurs when the males and females are both indexed by genotype, an unordered pair of alleles. Then Ω is the set of alleles, and A is the vector space of functions on Ω. The vector space of functions on genotypes is isomorphic with the symmetric square, so each parent factor is isomorphic with $\mathrm{sym}^2(A)$, or $\mathrm{sym}(A, B)$. To say the same thing in another way, each observation is associated with an ordered pair (female, male) of unordered pairs of alleles:

$$(\{r, s\}, \{i, j\}) \mapsto y_{r,s;i,j}.$$

In other words, y is an element of the tensor product space $A^{\otimes 4}$ in the subspace $(\mathrm{sym}^2(A))^{\otimes 2}$.

An example of a model that might be contemplated for a design of this type is the additive symmetric model

$$g(\mu_{r,s;i,j}) = \alpha_r + \alpha_s + \alpha_i + \alpha_j$$

in which all four alleles contribute equally and additively without interaction. This vector subspace is denoted by $\mathrm{symadd}(A, B, C, D)$. The single-locus Mendelian model is

$$\alpha_{ri} + \alpha_{sj} + \alpha_{rj} + \alpha_{si},$$

with $\alpha_{ij} = \alpha_{ji}$, an equally-weighted combination of the four sexually mixed genotypes. This is called the additive model with allelic interaction (Mather and Jinks, 1982). The symbolic expression for this subspace is

$$\text{symadd}\big(\text{sym}(A, C),\ \text{sym}(B, D),\ \text{sym}(A, D),\ \text{sym}(B, C)\big)$$

If it is thought that the sexes contribute unequally, an extended vector space containing functions of the form

$$\alpha_r + \alpha_s + \beta_i + \beta_j$$

or $\alpha_{rs} + \beta_{ij}$ might be tried. In practice, if the plants are self-sterile, any model may be restricted to the compatible pairs. All such models can be expressed symbolically using the operators described below.

2.6 *Further examples*

Other contexts in which similar designs arise include matrices of transition counts in Markov chains or more general processes with a discrete state space, allele-sharing of affected sib-pairs or triplets in the study of genetic diseases, cross-over designs with categorical responses, repeated measures (Agresti, 1997), social or occupational mobility tables, import-export analyses of international trade, and studies in which the observational unit is the nuclear family indexed by the political or religious affiliation of both parents (Cox 1972; Bloomfield 1974).

All of the models listed above determine vector subspaces not of the factorial type. Essential use is made of the fact that the factors are homologous. We now investigate the invariance properties of these vector spaces.

3. Factors, partitions and isomorphisms

Operationally, a factor is a list of levels, one level for each statistical unit. To say the same thing in another way, a factor is a function $\varphi: \mathcal{U} \to \Omega$ assigning to unit $u \in \mathcal{U}$ a factor level $\varphi(u) \in \Omega$. It is immaterial here whether φ has been assigned by randomization or whether the level is an intrinsic property of the unit. The inverse function φ^{-1} assigns to factor level $\omega \in \Omega$ the set of units that have that level:

$$\varphi^{-1}(\omega) = \{u \in \mathcal{U} : \varphi(u) = \omega\}.$$

If level ω does not occur, $\varphi^{-1}(\omega)$ is empty. In general $\varphi^{-1}(\Omega)$ is a partition of $\mathcal{U}$ into labelled blocks, possibly empty, one block of units for each factor level.

A factor A with levels Ω is here interpreted as the vector space

$$A = \mathcal{R}^{\Omega} = \{f : \Omega \to \mathcal{R}\}$$

of real-valued functions on the levels. By this definition, if B is another factor having the same set of levels, B is an identical copy of A. To each $f \in A$ there is a facsimile of f in B. This is a natural isomorphism between the two vector spaces.

Suppose now that $\varphi: \mathcal{U} \to \Omega$ is a list of factor levels in association with the units. Operationally, A may be realized as the subspace $\varphi^* \mathcal{R}^{\Omega} \subset \mathcal{R}^{\mathcal{U}}$ of functions on $\mathcal{U}$ that are constant on the blocks of the partition $\varphi^{-1}(\Omega)$. Likewise, if $\psi: \mathcal{U} \to \Omega$ is another list of factor levels, B may be realized as the subspace $\psi^* \mathcal{R}^\Omega \subset \mathcal{R}^\mathcal{U}$ of functions that are constant on the blocks of the partition $\psi^{-1}(\Omega)$. Furthermore, to each $v = f \circ \varphi$ in the subspace A, there corresponds a $v' = f \circ \psi$ in the subspace B. On the assumption that there are no empty blocks in either partition, this correspondence between $v \in A$ and $v' \in B$ is an invertible linear transformation on $\mathcal{R}^\mathcal{U}$.

Two subspaces are isomorphic in the vector-space sense if they have the same dimension. The subspaces A and B described above are isomorphic in a much stronger sense. They are isomorphic as representations of the symmetric group on Ω, as representations of the monoid on Ω, and as representations of the category of finite sets (McCullagh, 1997).

Given two vector subspaces $\mathcal{V}$ and $\mathcal{V}'$ that are isomorphic as representations, a new vector space $\mathcal{V}''$, isomorphic with $\mathcal{V}$ and $\mathcal{V}'$, can be constructed as follows. Let $\mathcal{V}' = g\mathcal{V}$ be the isomorphism. Given constants c, c', not both zero, define the set

$$\mathcal{V}'' = \{cv + c'g(v) : v \in \mathcal{V}\}.$$

If $c' = 0$, $\mathcal{V}''$ coincides with $\mathcal{V}$. Likewise, if $c = 0$, $\mathcal{V}''$ coincides with $\mathcal{V}'$. Otherwise, if $\mathcal{V}' \neq \mathcal{V}$, $\mathcal{V}''$ is a vector space distinct from $\mathcal{V}$ and $\mathcal{V}'$, but isomorphic with both. There is, however, one exception to this rule, which occurs only if $\mathcal{V} \cap \mathcal{V}' \neq 0$. It may happen that $cv + c'g(v)$ is identically zero on a subspace $\mathcal{V}_0 \subset \mathcal{V} \cap \mathcal{V}'$. The most common instance arises when g is the identity on $\mathcal{V}_0$ and $c' = -c$: at any rate, $g = -c/c'$, a scalar on $\mathcal{V}_0$. In general, the new subspace $\mathcal{V}''$ is isomorphic with the quotient spaces $\mathcal{V}/\mathcal{V}_0 \cong \mathcal{V}'/\mathcal{V}_0$.

It is necessary to define a computational operator that constructs a new vector space from a list of isomorphic vector spaces $\mathcal{V}_1 \cong \mathcal{V}_2 \cong \cdots$ in the manner described above. In general, there may be more than one natural isomorphism between two vector spaces: we assume here that the relevant isomorphism can be inferred from the context. Since this is an additive operation, we denote the operator by 'add' with syntax

$$\mathrm{add}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathrm{coeff} = c)$$

where c is a list of coefficients. Two special cases are sufficiently important and common in applications to merit special notation.

$$\begin{aligned} \mathrm{symadd}(\mathcal{V}_1, \mathcal{V}_2, \cdots) &= \mathrm{add}(\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathrm{coeff} = (1, 1, \ldots)) \\ \mathrm{altadd}(\mathcal{V}_1, \mathcal{V}_2) &= \mathrm{add}(\mathcal{V}_1, \mathcal{V}_2, \mathrm{coeff} = (1, -1)) \end{aligned}$$

These operators are also denoted by $\mathrm{sym}(\mathcal{V}_1 + \mathcal{V}_2 + \cdots)$ and $\mathrm{alt}(\mathcal{V}_1 + \mathcal{V}_2)$ respectively, but the displayed form is more suitable for computational purposes. The additive symmetric operator takes an arbitrary number of isomorphic vector spaces as arguments; the additive skew-symmetric operator takes two and only two arguments. Note that if A and B are homologous factors, the subspaces $\mathrm{altadd}(A, B)$ and $\mathrm{symadd}(A, B)$ in $A \otimes B$ are isomorphic with $A/1$ and A respectively.

The arguments of 'add' may be factors, but they may also be vector spaces that are not factors. Every factor corresponds to a labelled partition of the units. A compound factor such as $A.B$ also corresponds to a partition of the units, the greatest lower bound of the two partitions (Tjur 1984; Bailey 1991, 1996). However, there is no partition corresponding to the vector space $A + B$, the span of two factors. The following examples, in which

$A \cong B \cong C \cong D$ are homologous factors, serves to illustrate the range of possibilities. The levels of the four factors are denoted by r, s, t, u respectively.

Model-formula	Index notation	Dimension
$A + B$	$\alpha_r + \beta_s$	$2k-1$
$\mathrm{add}(A, B, (\cos\theta, \sin\theta))$	$\alpha_r \cos\theta + \alpha_s \sin\theta$	k^*
$\mathrm{symadd}(A, B, C)$	$\alpha_r + \alpha_s + \alpha_t$	k
$\mathrm{altadd}(A, B)$	$\alpha_r - \alpha_s$	$k-1$
$\mathrm{altadd}(A+B,\ B+C)$	$(\alpha_r + \beta_s) - (\alpha_s + \beta_t)$	$2k-2$
$\mathrm{symadd}(\mathrm{altadd}(A, B),\ \mathrm{altadd}(C, D))$	$\alpha_r - \alpha_s + \alpha_t - \alpha_u$	$k-1$
$\mathrm{altadd}(\mathrm{symadd}(A, B),\ C)$	$\alpha_r + \alpha_s - \alpha_t$	k
$\mathrm{altadd}(\mathrm{altadd}(A, B),\ \mathrm{altadd}(C, D))$	$\alpha_r - \alpha_s - \alpha_t + \alpha_u$	$k-1$
$\mathrm{symadd}(\mathrm{altadd}(A, B),\ \mathrm{altadd}(D, C))$	$\alpha_r - \alpha_s - \alpha_t + \alpha_u$	$k-1$
$\mathrm{altadd}(A.B,\ B.C)$	$\gamma_{rs} - \gamma_{st}$	$k(k-1)$

* For fixed θ. The dimension is $k-1$ if $\theta = 3\pi/4 \pmod{\pi}$.

In the right column, α, β are points in $\mathcal{R}^{\Omega}$, that is to say functions on the factor levels. Likewise, γ with two arguments or subscripts is a point in $A^{\otimes 2}$, or equivalently, in $A.B$, that is to say, a function on ordered pairs of factor levels.

It should be clear from inspection of the preceding list that two apparently distinct expressions may describe identical subspaces. For example $\mathrm{altadd}(A+B, B+C)$ is the same as $\mathrm{altadd}(A, B) + \mathrm{altadd}(B, C)$. Likewise $A+B$ is the same as $\mathrm{symadd}(A, B) + \mathrm{altadd}(A, B)$. Finally, although $\mathrm{altadd}(A, B) = \mathrm{altadd}(B, A)$, the two expressions

$$\mathrm{symadd}(\mathrm{altadd}(A, B),\ \mathrm{altadd}(C, D)) \quad \text{and} \quad \mathrm{symadd}(\mathrm{altadd}(A, B),\ \mathrm{altadd}(D, C))$$

give rise to distinct vector subspaces because of the ambiguity of the isomorphisms. The first 'symadd' operator takes $A \cong C$ and $B \cong D$ in order of appearance as the implied isomorphism: the second takes $A \cong D$ and $B \cong C$, also in order of appearance.

In the final line of the table, the two arguments of altadd are vector spaces whose intersection is equal to B, and the isomorphism is the identity on B. Consequently $\mathrm{altadd}(A.B,\ B.C)$ is a vector space isomorphic with $A.B/B$, of dimension $k^2 - k$. However, as the expression on line 7 in the table shows, the dimension of $\mathrm{altadd}(\mathcal{V}, \mathcal{V}')$ need not be equal to $\dim(\mathcal{V}) - \dim(\mathcal{V} \cap \mathcal{V}')$: the implied isomorphism does matter.

Expressions such as

$$\mathrm{symadd}(A + B,\ C), \quad \mathrm{symadd}(\mathrm{altadd}(A, B),\ C),$$

which have non-isomorphic subspaces where isomorphic subspaces are required, are forbiden. Note that $\mathrm{symadd}(A, B)$ is isomorphic with A, B, and hence with C: $\mathrm{altadd}(A, B)$ is isomorphic with $A/1$, and thus not isomorphic with any factor.

4. Further operators

In the sense that factorial models are complete, the following list of operators, together with those defined above, is complete for describing the relevant set of subspaces of $A \otimes B$ when $A \cong B$ are homologous. The list is not complete if there are three or more homologous factors.

$$\begin{aligned} A.B &= \{f : \Omega \times \Omega \to \mathcal{R}\} \\ \mathrm{sym}(A, B) &= \{f \in A.B : f(r, s) = f(s, r)\} \\ \mathrm{sym}^*(A, B) &= \{f \in \mathrm{sym}(A, B) : f(r, r) = 0\} \\ 1 &= \{f \in \mathrm{sym}(A, B) : f(r, s) = \mathrm{const}\} \\ \mathrm{alt}(A, B) &= \{f \in A.B : f(r, s) = -f(s, r)\} \end{aligned}$$

It should be emphasized here that the operator 'add' takes isomorphic vector spaces as its arguments. The operators listed above take only isomorphic *factors* as arguments, that is to say that the arguments must be vector spaces corresponding to partitions of the units. Now, $A.B$ and $\mathrm{sym}(A, B)$ are factors, so such expressions may occur as arguments where factors are required. For example, expressions such as $\mathrm{sym}(A, B).\,\mathrm{sym}(C, D)$, $\mathrm{sym}(A.B, B.C)$ and $\mathrm{alt}(\mathrm{sym}(A, B), \mathrm{sym}(B, C))$ are legitimate if $A \cong B \cong C$. By contrast, $\mathrm{sym}^*(A, B)$, $\mathrm{alt}(A, B)$, $\mathrm{add}(A, B, c)$ and $A + B$ return vector spaces that are not factors, so these cannot occur as arguments where factors are required.

Each of these operators has an obvious extension to three or more factors as arguments. For example,

$$\begin{aligned} \mathrm{sym}(A, B, C) &= \{f \in A.B.C : f(r, s, t) = f(t, r, s) = \cdots = f(r, t, s)\} \\ \mathrm{sym}^*(A, B, C) &= \{f \in \mathrm{sym}(A, B, C) : f(r, s, t) = 0 \quad \text{if } r = s,\ r = t,\ \text{or } s = t\} \\ \mathrm{alt}(A, B, C) &= \{f \in A.B.C : f(r, s, t) = f(t, r, s) = f(s, t, r) = -f(s, r, t)\} \end{aligned}$$

In other words, sym is completely symmetric under index permutation, sym^* is zero if there are duplicate indices, and alt is completely skew, taking one value for positive permutations and the negative value for negative permutations.

Some relationships among the subspaces are as follows. First, we have a decomposition into non-overlapping non-isomorphic subspaces:

$$\begin{aligned} A \otimes B &= \mathrm{sym}(A, B) \oplus \mathrm{alt}(A, B) \\ &= \mathrm{sym}^*(A, B) \oplus \mathrm{symadd}(A, B) \oplus \mathrm{alt}(A, B). \end{aligned}$$

These three subspaces are M-indecomposable, which is to say that they cannot be expressed as the direct sum of M-invariant subspaces. The first subspace is M-irreducible, but the second contains 1 and the third contains $\mathrm{altadd}(A, B)$. This is a complete M-decomposition of the symmetric and alternating subspaces.

In addition, the following subspaces, although not monoid-invariant, are sometimes useful in applied work:

$$\begin{aligned} \mathrm{diag}(A, B) &= \{f \in \mathrm{sym}(A, B) : f(r, s) = 0 \quad (r \neq s)\} \\ 1_d &= \{f \in \mathrm{diag}(A, B) : f(r, r) = f(s, s)\} \end{aligned}$$

The example that follows illustrates how these operators may be used to specify the model subspaces in complicated problems.

5. Transmission disequilibrium model

Consider a marker locus for which the set of alleles is $\Omega = \{M_1, \ldots, M_k\}$. Let Q_{ij} be the probability that an individual parent is of genotype $\{M_i, M_j\}$ and transmits M_i to his/her offspring. Ordinarily, $Q_{ij} = Q_{ji}$, each available allele being transmitted with probability one half. Suppose, however that the data (genotype of parent, transmitted genotype), are ascertained by sampling offspring who are affected by a particular disease D, suspected to have a genetic component. The relevant probability for data-analytic purposes is the conditional probability that the parent has genotype $\{M_i, M_j\}$ and transmits M_i, given that the offspring has the disease. In a fairly transparent notation,

$$\pi_{ij} = \mathrm{pr}(\{M_i, M_j\}, M_i \,|\, D) \propto \mathrm{pr}(D \,|\, M_i)\, Q_{ij}$$

under the assumption that only the transmitted allele can have any bearing on the probability of disease. If the marker locus and one disease-susceptibility locus are genetically linked on the same chromosome, and if the population is not in genetic equilibrium, it may happen that $\mathrm{pr}(D \,|\, M_i)$ depends on M_i. In the absence of linkage $\log \pi$ lies in $\mathrm{sym}(A, B)$. More generally,

$$\log \pi_{ij} = \log Q_{ij} + \alpha_i,$$

where $\alpha_i = \log \mathrm{pr}(D \,|\, M_i)$, which is quasi-symmetry, $\log \pi \in \mathrm{sym}(A, B) + A$. The transmission disequilibrium test is a test for symmetry given quasi-symmetry, by projection on to the quotient space $(\mathrm{sym}(A, B) + A)/\,\mathrm{sym}(A, B)$. This space is isomorphic with $A/1$ of dimension $k-1$. In more familiar statistical terminology, the test is based on the reduction in residual deviance caused by extending the model from $\mathrm{sym}(A, B)$ to $\mathrm{sym}(A, B) + A$.

A slightly more elaborate version of this argument is required if data are available in the form of three genotypes, one each for male parent, female parent and a single affected offspring. Under Mendelian rules of inheritance, the conclusion is that the log of the conditional probability of genotypes lies in the space

$$\mathrm{sym}(A, C) + \mathrm{sym}(A, B).\,\mathrm{sym}(C, D).$$

Slightly stronger, but often reasonable, assumptions lead to the subspaces

$$\mathrm{sym}(A, C) + \mathrm{sym}(\mathrm{sym}(A, B),\ \mathrm{sym}(C, D)) \quad \text{and} \quad \mathrm{sym}(A, C) + \mathrm{sym}^4(A, B, C, D).$$

In this notation, $\mathrm{sym}(A, B), \mathrm{sym}(C, D)$ are the factors generated by the parental genotypes, $\mathrm{sym}(A, C)$ is the factor whose levels are the offspring genotypes, and $\mathrm{sym}(C, D)$ is the factor corresponding to the untransmitted genotypes. In the absence of linkage, only the second term occurs in each model.

The preceding expressions are derived as follows. Let $y_{rs;ij}$ be the number of families (triples) for which the the male and female parental genotypes are $\{r, s\}$ and $\{i, j\}$ respectively, and the offspring genotype is $\{r, i\}$. As a function on Ω^4, y has no automatic symmetries. There is, admittedly, a slight problem to the extent that if both parents are heterozygous with the same genotype, only the sum $y_{rs;sr} + y_{sr;rs}$ is typically identifiable, in which case the observed total is allocated 50% to each cell. Mendel's rules assert that the corresponding probability $Q_{rs;ij}$ is doubly symmetric:

$$Q_{rs;ij} = Q_{sr;ij} = Q_{rs;ji} = Q_{sr;ji}.$$

This is the basis of the first model. The second model assumes also that the marker locus is not sex-linked, i.e. that sex and genotype are independent, $Q_{rs;ij} = Q_{ij;rs}$. The third model assumes complete symmetry, that is to say $Q_{rs;ij} = Q_{rj;is}$, so that the mixed genotypes are equally likely.

Regardless of the assumptions made concerning Q, the modulating factor arising from sampling affected offspring is

$$\mathrm{pr}(D \mid (r, i)) = \mathrm{pr}(D \mid (i, r))$$

on the assumption that the source of the allele is immaterial. In the preceding symbolic description of subspaces, A, B, C, D are factors whose levels are the four indices of y and Q in list order, so the levels of A and C are the paternal and maternal transnmitted alleles respectively. The expression $\log Q \in \mathrm{sym}(A, B).\,\mathrm{sym}(C, D)$ states that $\log Q$ is doubly symmetric: this space has dimension $k^2(k+1)^2/4$. The space $\mathrm{sym}(\mathrm{sym}(A, B), \mathrm{sym}(C, D)$ has dimension $k(k+1)(k^2+k+2)/8$. The completely symmetric subspace, $\mathrm{sym}^4(A, B, C, D)$, has dimension $k(k+1)(k+2)(k+3)/4!$. In all cases, the modulating factor $\log \mathrm{pr}(D \mid (r, i))$ lies in $\mathrm{sym}(A, C)$, although one might aim to sharpen the inferences by focusing on the additive and diagonal subspaces. The test for linkage is based on the quotient space projection $(\mathcal{V} + \mathrm{sym}(A, C))/\mathcal{V}$ where $\log Q \in \mathcal{V}$ represents the assumption made about symmetries of Q. In all three cases, this quotient space is isomorphic with $\mathrm{sym}(A, C)/1$ of dimension $k(k+1)/2 - 1$. If the design is incomplete, the actual dimension may be reduced.

Since $\mu_{rs;sr} = \mu_{sr;rs}$ for all models discussed above, allocation of the observed total equally to the two cells is in accord with the EM principle. All likelihood computations proceed as if both cells were observed independently, but the residual degrees of freedom must be reduced by one.

The relevant points so far as the present paper is concerned, are not the details of transmission disequilibrium models, but the method of specifying subspaces. To be useful, the symbolic description of subspaces must be capable of specifying various symmetries in a natural and succinct fashion.

Details of the transmission/disequilibrium test can be found in Spielman et al. (1993), Ewens and Spielman (1995) and Sethuraman and Speed (1997).

6. Computer implementation

A partial implementation of these operators in 'S' is described by Glonek and McCullagh (1997).

Acknowledgements: T.P. Speed brought the transmission disequilibrium test to my attention. Gary Glonek showed that the operators can be implemented in 'S'. Burt Totaro gave extensive help with the representation theory.

References

Agresti, A. (1983) Testing marginal homogeneity for ordinal categorical variables. *Biometrics* **39**, 505–510.

Agresti, A. (1997) A model for repeated measurements of a multivariate binary response. *J. Am. Statist. Assoc.* **92**, 315–21.

Bailey, R.A, (1991) Strata for randomized experiments (with discussion). *J. R. Statist. Soc.* B **53**, 27–78.

Bailey, R.A. (1996) Orthogonal partitions in designed experiments. *Design codes and cryptography* **8**, 45–77.

Bloomfield, P. (1974) Linear transformations for multivariate binary data. *Biometrics* **30**, 609–617.

Caussinus, H. (1965) Contribution à l'analyse statistique des tableaux de correlation. *Ann. Fac. Sci. Univ. Toulouse* **29**, 77–182.

Cox, D.R. (1972) The analysis of multivariate binary data. *Appl. Statist.* **21**, 113–120.

Diaconis, P. (1988) *Group Representations in Probability and Statistics.* IMS Lecture Notes **11**, Hayward, CA.

Diaconis, P. (1989) A generalization of spectral analysis with application to ranked data. *Ann. Statist.* **17**, 949–79.

Ewens, W.J. and Spielman, R.S. (1995) The transmission/disequilibrium test: history, subdivision and admixture. *Amer. J. Human Genetics* **57**, 455–64.

Glonek, G. and McCullagh, P. (1997) An extended model syntax for designs involving homologous factors. Technical Report 448.

Griffing, B. (1956) Concept of general and specific combining ability in relation to diallel crossing systems. *Austral. J. Biol. Sci.* **9**, 463–93.

Mather, K. and Jinks, J.L. (1982) *Biometrical Genetics.* Chapman and Hall, London.

McCullagh, P. (1982) Some applications of quasisymmetry. *Biometrika* **69**, 303–8.

McCullagh, P. (1997) Invariance and linear models. Technical Report 450

Nelder, J.A. (1977) A reformulation of linear models, (with discussion). *J. R. Statist. Soc.* A **140**, 48–77.

Ott, J. (1989) Statistical properties of the haplotype relative risk *Genetic Epidemiology* **6**, 127–30.

Sethuraman, B. and Speed, T. (1997) Transmission disequilibrium test via marginal homogeneity. Unpublished manuscript.

Spielman, R.S., McGinnis, R.E. and Ewens, W.J. (1993) Transmission test for linkage disequilibrium. *Amer. J. Human Genetics* **52**, 506–16.

Stewart, J.Q. (1948) Demographic gravitation: Evidence and applications. *Sociometry* **1**, 31–58.

Stigler, S. (1994) Citation patterns in the journals of statistics and probability. *Statistical Science* **9**, 94–108.

Tjur, T. (1984) Analysis of variance models in orthogonal designs (with discussion). *Int. Statist. Rev.* **52**, 33–81.

Upton, G.J.G. (1985) Modelling cross-tabulated regional data. In *Measuring the Unmeasurable* Ed. P. Nijkamp. NATO Scientific Publications, 1985, 197–218.

Wilkinson, G.N. and Rogers, C.E. (1973) Symbolic description of factorial models for analysis of variance. *Appl. Statist.* **22**, 392–9.

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