INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

SMR.780 - 55

FOURTH AUTUMN COURSE ON MATHEMATICAL ECOLOGY

(24 October - 11 November 1994)

"Affinity in Paired Event Probability"

Carlos Castillo-Chavez
Biometrics Unit
Cornell University
Ithaca, NY 14853-7801
U.S.A.

These are preliminary lecture notes, intended only for distribution to participants.

To appear in
Math. Biosci.

TECHNICAL REPORT
94-40

AFFINITY IN PAIRED EVENT PROBABILITY

BY

Stephen P. Blythe, Stavros Busenberg
and
Carlos Castillo-Chavez

JUNE 1994

Supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University, Contract DAAL03-91-C-0027

AFFINITY IN PAIRED EVENT PROBABILITY[1]

Stephen P. Blythe[2], Stavros Busenberg[3], and Carlos Castillo-Chavez[4]

Abstract

It is shown that a general parametric functional generates the conditional and joint probabilities of event pairs when the order within paired events is irrelevant. The parameters represent affinities or associations between single events. If the marginal probabilities of the single events are known, then these parameters specify a hypersurface on which all the joint probabilities of event pairs must lie. Examples are presented and applications in probability, ecology, epidemiology, genetics, and distribution theory are offered.

[1]An advanced draft was completed on March 2, 1993, a month before the death of Stavros Busenberg.

[2]Department of Statistics and Modelling Science, Strathclyde University, Livingston Tower, Glasgow G1 1XH, Scotland.

[3] Department of Mathematics, Harvey Mudd College, Claremont, CA 91711, USA.

[4] Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853, USA.

1 Introduction

In the epidemiology of sexually transmitted diseases, the problem of modeling arbitrary mixing patterns of distinct population sub-groups (differentiated, for example, by age, sex, number of partners, etc.) has received a great deal of attention in the last few years, partly due to the implications of the AIDS pandemic. Furthermore, the closely connected problem of pairing–partnership formation–has also received considerable attention (see Dietz and Hadeler 1988; Castillo-Chavez and Busenberg 1991; Castillo-Chavez, Busenberg, and Gerow 1991). Recently, a general solution to the mixing or pairing problem has been found (Busenberg and Castillo-Chavez, 1989, 1991). One of the forms of this solution is a parametrized multiplicative perturbation of random mixing, the nonrandomness arising from the preference or affinity between particular subgroups (see Blythe *et al.* 1991). Because sexual mixing is essentially a paired event process, it is possible to express the problem and the solution in general probabilistic terms. This re-formulation of the problem of mixing extends the applicability of the original mixing problem solution. The aim of this paper is to present this extension and give examples that show how it can be applied to a variety of paired event problems.

In this article we show that the axiomatic framework underlying the theorems of Busenberg and Castillo-Chavez (1989, 1991) is equivalent to specifying a process where two elementary events occur as an "event pair," with the order of elementary events within a pair being immaterial. Hence, we show that a parameterized representation of all such processes may be stated explicitly. We use a number of examples to illustrate the form and structure of the paired event representation, and indicate a few areas where it is applicable. We conclude with a brief discussion of these applications and indicate directions for future work.

2 Paired Event Probability

Let $S = \{1,2,\cdots,n\}$ be an index set, and $\Omega = \{E_i : i \in S\}$ be the finite sample space of disjoint elementary events E_i. We use the following notation:

$\pi_{ij} \equiv \Pr\{(E_i,E_j)\ (i,j) \in S^2,\ (E_i, E_j) \in \Omega^2\}$: the probabilities of event pairs.

$\rho_i \equiv \pi_{i.} \equiv \sum_{j=1}^{n} \pi_{ij} \equiv \Pr\{E_i\},\ i \in S,\ E_i \in \Omega$: the marginal probabilities for events.

$p_{B|A} \equiv \Pr\{(A,B)|B\}$: the conditional probabilities of paired events.

$p_{ij} \equiv \pi_{ij|i} \equiv \Pr\{(E_i, E_j)|\ E_i\}$: the conditional probabilities of paired events.

From the above definitions we have that:

$$\rho_i \in [0,\ 1],\ i \in S\ , \tag{1}$$

$$\sum_{i \in S} \rho_i = 1\ , \tag{2}$$

and

$$\text{if } \rho_i = 0 \text{ or } \rho_j = 0 \text{ then } p_{ij} = p_{ji} = \pi_{ij} = \pi_{ji} = 0\ , \tag{3}$$

reflecting the fact that the ρ are probabilities, and that zero probability events cannot occur in event pairs. We impose the *nontrivial* restriction

$$\pi_{ij} = \pi_{ji},\ (i,j) \in S^2\ , \tag{4}$$

which specifies the class of problems to be addressed: those where the probabilities of paired events do not depend on the order of the pair. We are assuming that the joint distribution be symmetric but that (i,j) and (j,i) be considered distinct events. Therefore, despite (4) it is possible to have $p_{ij} \neq p_{ji}$, and this will indeed be the case in most of the situations that we shall examine.

We note the following immediate consequences of the above restrictions and definitions:

i. $p_{ij} \geq 0$,

ii. $\sum_j p_{ij} = 1$, if $\rho_j \neq 0$,

iii. $\rho_i p_{ij} = \rho_j p_{ji}$,

iv. If $\rho_i \wedge \rho_j = 0$ then $p_{ij} = 0$.

We then have the following representation result.

Theorem 1 *Let S' be the subset of S where $\rho_i \neq 0$. For any given marginal distribution of elementary events ρ, there exists an $n \times n$ symmetric matrix of constants ϕ, with $\phi_{ij} = \phi_{ji} \geq 0$, such that*

$$\pi_{ij} = \rho_i \rho_j \left[\frac{R_i R_j}{\sum_{m \in S} \rho_m R_m} + \phi_{ij} \right],\ (i,j)\ \in\ S' \times S' \tag{5}$$

where

$$R_i = 1 - \sum_{m \in S} \rho_m \phi_{im},\ i \in S' . \tag{6}$$

The ϕ are constrained by $R_i \geq 0$, and the first term in the bracket in (5) is taken to be zero when all the R_i are zero.

Proof Conditions (i)–(iv) constitute the axioms of Busenberg and Castillo-Chavez (1989, 1991), and so their representation theorem (Theorem 4.4 in their 1991 paper) for the matrix $(p_{ij}: (i,j) \in S' \times S')$ holds, with ϕ satisfying the above hypotheses and R_i given by (6). Thus

$$p_{ij} = \rho_j \left[\frac{R_i R_j}{\sum_{m \in S} \rho_m R_m} + \phi_{ji} \right]. \tag{7}$$

The expression (5) for π_{ij} is obtained by using (7) and Bayes' theorem. □

Remarks:

(a) The parameters ϕ are a measure of affinity of each kind of event for every other kind. If $\phi_{ij} = x$, $x \in [0,1]$ for all $(i,j) \in S'$, then (5) reduces to

$$\pi_{ij} = \rho_i \rho_j , \tag{8}$$

which is the familiar result for independent events. Different choice of ϕ—a total of $n(n+1)/2$ possibilities—produce all the possible joint distributions of events in pairs, where the probabilities of paired events are independent of the order.

(b) Knowledge of $\{\rho_i: i \in S'\}$; the assumption $\pi_{ij} = \pi_{ji}$, that is, property (iii); and property (ii) imply the existence of only $n(n-1)/2$ independent values in the matrix $(p_{ij}: (i,j) \in S' \times S')$. This matrix according to Equation (7) requires $n(n+1)/2$ "independent" ϕ-values.

(c) The ϕ-values must satisfy the conditions $R_i \geq 0$ or, equivalently, the set of inequalities

$$0 \leq \sum_{m \in S} \rho_m \phi_{im} \leq 1,\ i \in S' . \tag{9}$$

(d) Let a be a real number then if $\phi_{ij} = a$, $\forall\ i,j \in S$ the set of inequalities given by (9) imply that $a \in [0,1]$. In fact, the endpoint values are achieved when $a=0$ and $a=1$. Sets of non-negative constant values for the elements of the ϕ-matrix that include ϕ-values larger than 1, for which the constraints imposed by (9) are not violated, are possible.

(e) Bahadur (1961) and Lazarfeld (1956) have looked at the representation of joint distributions of responses to n dichotomous items. He has constructed representation theorems for these distributions that are given as multiplicative perturbations of the joint distribution that would be obtained if these variables were independent. Bahadur 's representation theorem is not in terms of an affinity matrix

but rather in terms of a polinomial expansions that take advantage of the fact that the processes that he modeled were discrete (binomila or multinomial processes). Our representation theorem collapses to Bahadur's in these special situations. However, our representation theorem can also handled continuos distributions. In addition, Bahadur's conditions so that the pairing axioms are satisfied are very difficult to check. Furthermore, his representation theorem does not provide a clear way of constructing large families of parametric distributions for correlated paired events.

3 Examples and Applications

In this section we present several examples demonstrating the applicability of the above theorem. We start with a very simple situation (we will return to this example later on).

Example 1: Color-Bias in Card Games

In a caricature of the card game patience (solitaire), a game consists of shuffling the deck, and dealing out the top two cards. If these are the same color the player wins, otherwise she loses. When the deck is fair, then two reds occur with probability 0.25, two blacks with probability 0.25, and two mismatched cards with probability 0.5. For an *unfair* deck, where cards have some tendency to remain together through the shuffle, on the basis of color alone, the outcomes will differ from the random result. We may use the method of the previous section, the "ϕ-method," to examine the problem of affinities between cards of different colors.

Here we have two events: E_1 = "red card", and E_2 = "black card", with $\rho_1 = \rho_2 = \frac{1}{2}$. If we write

$$\phi = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \tag{10}$$

and if we define

$$\alpha = \frac{1}{4}\,\frac{4(1-b)+b^2-ac}{4-(a+2b+c)}, \tag{11}$$

we obtain from (5) the following form for π:

$$\pi = \begin{pmatrix} \alpha & \frac{1}{2}-\alpha \\ \frac{1}{2}-\alpha & \alpha \end{pmatrix}. \tag{12}$$

Note that π contains only one undetermined parameter α which, if known, places a restriction on the four parameters that are present in ϕ, but does not determine them uniquely. We have

$$\text{if } \phi = \begin{pmatrix} 2 & 0 \\ 0 & c \end{pmatrix} \text{ then } \pi = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}, \tag{13}$$

$$\text{if } \phi = \begin{pmatrix} a & a \\ a & a \end{pmatrix} \text{ then } \pi = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \tag{14}$$

$$\text{if } \phi = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix} \text{ then } \pi = \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix}. \tag{15}$$

This example gives an illustration of the value of the ϕ-method. Consider the ϕ values as a measure of color-for-color affinity among cards, brought about by the shuffling process; $\phi_{ij} = a$, $a \in [0,1]$, implies no affinity, while $\phi_{ij} = 2\delta_{ij}$ (δ_{ij} the Kronecker delta) in this example implies maximum affinity. If all affinities are equal ($\phi_{ij} = a$, $a \in [0,1]$), the result is a random pattern (14). If just cards of one color have maximum mutual affinity (that is, R_i=0 for either i =1 or i=2), and there are zero inter-color affinities, then cards of that color are always found together, so that color separation takes place, and the result is a guaranteed win, regardless of the self-affinity of the other color (13). To get a guaranteed loss [no same-color pairs (15)], the inter-color affinities must be maximum (that is, R_i=0 for i=1,2), and both self-affinities must be zero. This last fact is due to the required symmetry of the ϕ-matrix. Furthermore, note that we could not have $\phi_{ij} = 2 \; \forall$ i, j $\in S'$.

"Affinity" here really means that the shuffling process favors or disfavors like-with-like color matches, making cards adjacent with a probability greater than random. The matrix ϕ thus measures bias in the shuffle, and provides a way of quantifying the degree of certain kinds of non-randomness. It is interesting to note that bias (or cheating), which causes a *particular* color always to win, involves changing the marginals, setting say $\rho_1 = 0$ for black to win with probability one.

It is possible to use the ϕ-method in the analysis of a variety of paired event experiments, perhaps in the sense discussed by Mielke and Siddiqui (1965), where independence is normally "expected," but where outside influences sometimes correlate occurrences (in their case, the presence of atmospheric pollutants caused nonrandom temporal clustering of asthmatic attacks among three subjects).

Example 2: Ecological Association in Plant Communities

A problem of long-standing interest in plant ecology is the degree to which species tend to be found together in particular habitats (see Ludwig and Reynolds, 1988, the source for much of the following background information). Legendre and Legendre (1983) refer to this kind of study as *R-mode analysis*. One particular set of questions concerns inter-specific association (also referred to in the literature as affinity), arising because (a) both species select or avoid the same kinds of habitat, (b) both species have approximately the same requirements, or (c) there is a genuine affinity or disaffinity between them (Hubalek 1973, Ludwig and Reynolds, 1988). Ludwig and Reynolds (*op. cit.*, Chapter 11) discuss this at length, and present many practical schemes for detecting and classifying associations. An assortment of indices, measures, and statistical tests have been developed for studying associations, some of which work quite effectively when tested on known, artificial data (*op. cit.*). Both presence-absence indices (association) and those based on abundance (covariation) have

been developed. All these measures are essentially *ad hoc* in nature. The paired event affinity structure of (5) may provide a useful framework for investigating association. The key result which remains to be obtained is the elaboration of precise experimental protocols for sampling. We speculate that an experimental design capturing the essential unordered event pair characteristic of the method would involve two linked points in an appropriate sampling unit: perhaps by looking at the two nearest plants in each of a series of randomly chosen points from a unit, and replicating over many units to get the multiplicity of sampled joint distributions required to specify uniquely the elements of ϕ.

If data on π can be obtained for several (say M) replicates, then we can estimate the ϕ-values directly using, for example, the method of least squares. There are $n(n+1)/2$ independent ϕ-values, but only $n(n-1)/2$ independent values in a p or π matrix, so that we require at least $n+1$ samples to fit the ϕ.

As a preliminary example of how the ϕ-method might be applied, we choose a data set where we can approximately (but only approximately) extract the information we need to use the method, namely that of Krebs (1978, p. 375 *et seq.*). An ecology class looked at the association of two species of grasses on an area of sand dunes on the southern shores of Lake Michigan, performed both quadrant presence-absence and nearest-neighbor analyses. The species were *Andropogon scoparius* (Species 1 here) and *Ammophilia breviligulata* (Species 2 here). The data are not ideal for our purposes, but we can extract two rough estimates of π, one from the presence-absence data, and one from the nearest-neighbor data. With E_1 = "Species 1 found" and E_2 = "Species 2 found" and using the notation $\pi^{(k)}$ for the k^{th} data matrix (with M = 2), we have

$$\pi^{(1)} = \begin{pmatrix} 0.3615 & 0.0308 \\ 0.0308 & 0.5769 \end{pmatrix}, \quad \text{hence} \quad \rho^{(1)} = \begin{pmatrix} 0.3923 \\ 0.6077 \end{pmatrix}, \tag{16a}$$

and

$$\pi^{(2)} = \begin{pmatrix} 0.3214 & 0.0863 \\ 0.0863 & 0.5060 \end{pmatrix}, \quad \text{hence} \quad \rho^{(2)} = \begin{pmatrix} 0.4077 \\ 0.5923 \end{pmatrix}, \tag{16b}$$

Thus we have two estimates of π for the dune grasses. Now, note that in this case where $n = 2$ only two elements in π, for each k, are independent, but that there are three independent ϕ values. This means that we cannot uniquely specify the ϕ values from the available data; however, we can test the fit of two contradictory hypotheses, namely, that (A): Intra-species affinity dominates, and (B): inter-species affinity dominates. We can express these hypotheses using test matrices for ϕ,

$$\mathfrak{H}_A\colon\ \phi = \begin{pmatrix} a & 0 \\ 0 & a \end{pmatrix}, \tag{17}$$

and

$$\mathfrak{H}_B\colon\ \phi = \begin{pmatrix} 0 & b \\ b & 0 \end{pmatrix}. \tag{18}$$

Of course, many other one-parameter families may be tested (e.g., all distinct combinations of 0, 1, and a parameter); the point here is to show that it can be done, not to perform an exhaustive analysis based on these limited data. We estimate the parameters in (17) or (18) using the obvious scheme of minimizing

$$S(\phi) = \left[\frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{n-1}\sum_{j=i}^{n}\left(\pi_{ij}^{(k)} - \hat{\pi}_{ij}\right)^2\right]^{1/2}, \tag{19}$$

where $\hat{\pi}_{ij}$ is the optimal value based on the data and which we try to match by using (5) and (6) with ϕ given in (17) or (18). The outer sum is over replicates, the next over rows in π, and the inner sum over columns; there are $M\left[\frac{n(n+1)}{2}-1\right]$ terms, equaling the number of independent π_{ij}. Using the two-parameter representation

$$\hat{\pi} = \begin{pmatrix} \alpha & \beta \\ \beta & 1-\alpha-2\beta \end{pmatrix} \tag{20}$$

which follows from Theorem 1, we obtain a positive definite quadratic form $Q(\alpha, \beta) = S^2(\phi)$ which has a unique minimum when $\alpha = 0.34145$ and $\beta = 0.05855$, yielding $S(\phi) = 0.03423554$ and

$$\hat{\pi} = \begin{pmatrix} 0.34145 & 0.05855 \\ 0.05855 & 0.54145 \end{pmatrix}, \qquad \begin{matrix} \hat{\rho}_1 = 0.4, \\ \hat{\rho}_2 = 0.6. \end{matrix} \tag{21}$$

In order to test the two hypotheses $\mathcal{H}_A$ and $\mathcal{H}_B$ we use the test of whether or not the corresponding ϕ can attain the optimum values of $\hat{\pi}$ given by (21) while remaining consistent with the restrictions of Theorem 1. In the case of $\mathcal{H}_A$ it is easily seen from (5), (6), and (17) that this will happen only if the parameter α satisfies the quadratic equation

$$\hat{\pi}_{11} = \hat{\rho}_1 - \frac{\hat{\rho}_1\hat{\rho}_2(1-a+\hat{\rho}_1\hat{\rho}_2 a^2)}{1-(\hat{\rho}_1^2+\hat{\rho}_2^2)a}.$$

This yields two positive values for a, namely $a = 1.420609$ and $a = 2.217482$, the second of which is inadmissible since it leads to a negative value for R_2. Thus, $\mathcal{H}_A$ with $a = 1.420609$ satisfies the test that we have adopted. As for $\mathcal{H}_B$, it is seen from (5), (6), and (18) that it can be admissible according to our test only if b satisfies the quadratic equation

$$\hat{\pi}_{11} = \hat{\rho}_1^2\,\frac{(1-\hat{\rho}_2 b)^2}{1-2\hat{\rho}_1\hat{\rho}_2 b},$$

which yields the values $b = -1.547602$ and $b = 2.035519$. Both of these are inadmissible, the first because it violates the positivity of ϕ_{ij}, and the second because it yields a negative value of R_1. Thus, $\mathcal{H}_B$ is rejected and $\mathcal{H}_A$ is accepted. We can thus be reasonably sure that the species disassociate, a conclusion that agrees with those of the standard tests described by Krebs (1978).

It is also interesting to compare hypothesis $\mathcal{H}_A$ with the assumption that the grasses mix randomly, that is, $\phi_{ij} = k$, $\forall\, i,j \in S'$, $k \in [0,1]$. Assuming ρ_i to have the optimal values $\hat{\rho}_i$ given by

(21), random mixing yields

$$\pi = \begin{pmatrix} 0.16 & 0.24 \\ 0.24 & 0.36 \end{pmatrix},$$

and a value of $S(\phi)$ equal to 0.258827, far from the optimal value 0.03423554 attained by $\mathcal{H}_A$ with the choice a = 1.420609. Of course, the fact that $\mathcal{H}_A$ satisfies this criterion does not mean that it gives the correct affinity matrix, since as we have discussed above, the solution for ϕ is not unique. However, based on the available data, $\mathcal{H}_A$ satisfies the test criterion of yielding the optimum $\hat{\pi}$ in (21) and minimizing $S(\phi)$ given by (19).

Example 3: Single-Sex Mixing Models

Example 3a. In its original form, the result of Busenberg and Castillo-Chavez (1989, 1991) deals with sexual contacts (i.e., partners) per unit time in a population comprised of n groups, in the i^{th} of which there are $T_i(t)$ individuals, at time t, with an average number of partners per unit time, or "risk" level, equal to $C_i(t)$. A valid description of the mixing process is produced by an $n \times n$ matrix of probabilities p(t) where $p_{ij}(t)$ is the probability that an individual in group i has a partner in group j, at time t (given that it paired). Constraints (i) and (ii) of Section 2 simply make p a stochastic matrix, (iii) enforces conservation of the number of new pairings per unit time between individuals in groups, and (iv) says that individuals in momentarily empty or inactive groups cannot have partners.

The problem of developing deterministic p(t), functions of C(t) and T(t) only, which satisfy the axiomatic constraints, has become one of extreme importance in modeling the epidemiology of sexually transmitted diseases (STDs), and in particular human immunodeficiency virus HIV, the causative agent of AIDS. Until recently, the only known solution was random or "proportionate" mixing (e.g., Barbour, 1978; Nold, 1980; Hethcote and Yorke, 1984; Anderson and May, 1984; Dietz and Schenzle, 1985; Anderson *et al.*, 1987; Blythe and Anderson, 1989) where $p_{ij} = \rho_j$, for all i, and with the marginals, in this context, being given by

$$\rho_i(t) = \frac{C_j(t)T_j(t)}{\sum_{k=1}^{n} C_k(t)T_k(t)}.$$

The representation theorem states that any p which is a solution to the problem specified by constraints (i)–(iv) of Section 2 may be written in the form given by (5). The parameters ϕ_{ij}, which may in this context be functions of elements of the matrix $(\rho_{ij} \equiv \rho_i(t))$, provide a measure of *mutual preference*, or *affinity* for sexual partners between pairs of groups (see Blythe and Castillo-Chavez, 1989; Busenberg and Castillo-Chavez, 1989, 1991; Castillo-Chavez and Blythe, 1989; and Castillo-Chavez *et al.*, 1991). A constant ϕ-matrix implies that the preference structure in the population remains unchanged, but it is important to note that the values of p will change with time as a result

of the dynamics of the model changing the sizes of the mixing subpopulations, $T_i(t)$; i.e., time dependent changes in the set of mixing probabilities $p_{ij}(t)$ are not necessarily an indication of change in behavior. Even where all of the n groups in the population are identical (i.e., all the $C_i(t)$ are equal, and all the $T_i(t)$ are equal, so that $p_j(t) = 1/n$ for all j), the preference structure ϕ can produce nonrandom mixing in p (*cf.* Blythe *et al.*, 1991).

Example 3b. Another area where sexual mixing models can be of some importance is population genetics. As a first approximation at applying the generalized mixing framework in this area, consider the case of a recessive gene at a single locus, with no population regulation or frequency dependence (Blythe *et al.*, 1991). A good example (Crow, 1986, pp. 50-53) is redheadedness in a human population. We divide the population into three groups: Group 1, homozygous (AA) individuals who do not carry the "red" allele; Group 2, heterozygotes (Aa) who carry but do not express the allele; and Group 3, heterozygotes (aa) who express red hair. We assume that the fractions of offspring born from the six possible crosses follow the standard elementary random pattern (see *op. cit.*).

Clearly no one can distinguish AA from Aa phenotypes, so the only reasonable form for the preference or affinity matrix is

$$\phi = \begin{pmatrix} \alpha & \alpha & \beta \\ \alpha & \alpha & \beta \\ \beta & \beta & \gamma \end{pmatrix}, \tag{24}$$

where $0 \leq \alpha, \beta, \gamma \leq 1$ are constants. Individuals in groups 1 and 2 have the same preferences for aa versus non-aa (β versus α), and group 3 individuals have preferences β and γ for non-aa and aa individuals, respectively. Assortative mating is represented by $\beta = \alpha = 0$, and $\gamma = r/\rho_3(t)$ where $r < 1$; i.e., a fixed fraction of the population of aa individuals mate among themselves, regardless of group population sizes.

In Blythe *et al.* (1991), it is shown how a simple model may be derived under the following standard simplifying assumptions: (a) every individual in generation t has just one partner; (b) the unit of time is the generation; (c) individuals from generation t are not counted in generation $t+1$; and (d) all matings produce 2ζ offspring ($\zeta > 0$). Then, $C_i(t) = 1$, for all i and t, and

$$\rho_i(t) = \frac{T_i(t)}{T(t)} = x_i(t) \text{ for all i, and } T(t) = \sum_{k=1}^{n} T_k(t)\,, \tag{25}$$

where $T(t)$ is the total population, and $x_i(t)$ is the proportion of group i in the population, in generation t. Note that $x_1(t) + x_2(t) + x_3(t) = 1$ and $\frac{1}{2}x_2(t) + x_3(t) = q(t)$—the a allele frequency.

Now, for convenience write $Z_t = x_3(t)$ as the proportion of aa individuals in the population in generation t. Then we have the recurrence relation

$$Z_{t+1} = f(Z_t) + \frac{g(Z_t)^2}{h(Z_t)}, \qquad Z_0 < q, \tag{26}$$

where

$$f(Z) = aq^2 - 2(\alpha-\beta)qZ + (\alpha-2\beta+\gamma)Z^2$$
$$g(Z) = (1-\alpha)q + (\alpha-\beta)(1+q)Z - (\alpha-2\beta+\gamma)Z^2$$

and

$$h(Z) = (1-\alpha) + (\alpha-\beta)Z - (\alpha-2\beta+\gamma)Z^2 .$$

Blythe *et al.* (1991) consider some of the properties of this map, and suggest directions in for future work in population genetics using this formalism.

Example 4: Joint Distributions

Equation (5) seems also to be a new result in the theory of distributions. Consider X and Y, two discrete random variables, with joint density function $f_{X,Y}(\cdot,\cdot)$ and marginal density functions $f_X(\cdot)$ and $f_Y(\cdot)$, respectively. If the joint density function is jointly symmetric, i.e., $f_{X,Y}(u, v) = f_{X,Y}(v, u)$ for all (X, Y) in the appropriate space, then clearly $f_X(u) = f_Y(u)$, for all u, so X and Y have the same function for their marginals. But these marginals are just ρ_u, and $\pi_{uv} \equiv f_{X,Y}(u, v)$, so that (5) gives us a representation of all jointly symmetric discrete joint density functions.

In fact, there is nothing new in the use of infinite families of joint distributions: for example, Mood *et al.* (1974, p. 142) use one to illustrate the fact that knowing the marginals does not necessarily mean that we know the joint distribution. Using discrete random variables, and a symmetric density function, we may write Mood *et al.*'s (1974) example as

$$\pi_{ij} = \rho_i\rho_j\left[1 + \nu(\Upsilon_i - Q)(\Upsilon_j - Q)\right], \tag{27}$$

where $-1 \le \nu \le +1$ is an arbitrary parameter,

$$\Upsilon_i \equiv \sum_{k=1}^{i} \rho_k , \tag{28}$$

is the cumulative distribution of the marginals, and

$$Q \equiv \sum_{m=1}^{n} \rho_k \Upsilon_k \tag{29}$$

(with continuous variables, $Q = \frac{1}{2}$). This example was originally due to D. Morgenstern (Plackett, 1965), and has been extended and studied by, e.g., Farlie (1960) and Gumbel (1960). Clearly, we always have the same marginals, ρ, regardless of the value of ν, and of course the constraints (i)–(iv) of Section 2 are satisfied. This means, incidentally, that the p_{ij} deriving from (27) could be used as a mixing function in STD modeling, as in Example 3 above. What the new representation result says is that all such examples may be expressed in the form of (5), it being required only to find the appropriate ϕ. It is easy to show that in Mood *et al.*'s (1974) example, the choice

$$\phi_{ij} = \frac{\nu}{1+\nu Q^2}\Upsilon_i\Upsilon_j \tag{30}$$

in (5) leads after some algebra to the family (27) as a special subclass of π, for $0 \leq \nu \leq 1$. For the negative – and perfectly permissible – range of ν, (30) leads to negative ϕ_{ij}, which is not permissible under the representation theorem. A set of values of ϕ which are nonnegative may be obtained, however, by using

$$\phi_{ij} = 1 + \nu(\Upsilon_i - Q)(\Upsilon_j - Q) \,, \tag{31}$$

which works, with nonnegative ϕ, for the full range of ν. In this case, all the $R_i = 0$, and (5) reduces to $\pi_{ij} = \rho_i \rho_j \phi_{ij}$. The relaxation of the assumption that the ϕ are strictly constants is important only if the marginals change with time, a point discussed in Blythe *et al.* (1991).

Remark

The above example matches that of Bhadur (1961) when he formulates his representation theorem for the joint distribution of correlated binomial variables. The above example also makes the point that the relationship between ϕ and π is not unique (which was, after all, essentially the point that Mood *et al.* (1974) were making). However, because (5) is a representation of all π, this nonuniqueness need no longer be regarded as something of a *bête noire*, as we show in the following example.

The extension to continuous variables is trivial, regaining the original formulation of Busenberg and Castillo-Chavez (1989, 1991). Consider X and Y, two jointly continuous random variables, with jointly symmetric density function $\pi(x, y)$, and hence marginal density functions $\rho(x)$ and $\rho(y)$. Then π is related to ρ by an equation exactly analogous to (5), with $R(\cdot)$ being now defined by an integral rather than a summation, and the denominator likewise. For example, if $\rho(x) = 1/h$ for $x \in [0, h]$ and zero elsewhere, i.e., uniformly distributed marginals, and say

$$\phi(x, y) = \frac{1}{2h}(x + y), \qquad (x, y) \in [0, h] \,, \tag{32}$$

then we directly have that

$$\pi(x, y) = \frac{1}{16h^2}\left[2 + 7\frac{xy}{h^2} + \left(4 - \frac{x}{h}\right)\left(4 - \frac{y}{h}\right)\right], \qquad (x, y) \in [0, h] \,. \tag{33}$$

This is a convenient way of generating joint density functions with a specified "affinity" structure $\phi(x, y)$, but the reverse process of finding $\phi(x, y)$ such that a given $\pi(x, y)$ is obtained is often difficult. For example, say $\pi(x, y)$ is the bivariate normal distribution

$$\pi(x, y) = \frac{1}{2\pi\sigma^2\sqrt{1-\xi^2}}\, e^{-\frac{1}{2\sigma^2(1-\xi^2)}\left[(x-\mu)^2 - 2\xi(x-\mu)(y-\mu) + (y-\mu)^2\right]} \,, \tag{34}$$

for $(x, y) \in (-\infty, +\infty)$, with μ and σ the mean and variance, respectively, of the marginal $\rho(\cdot)$, and ξ the correlation coefficient of X and Y. Clearly,

$$\rho(z) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(z-\mu)^2}{2\sigma^2}} \,. \tag{35}$$

To find suitable $\phi(x, y)$ we have to equate

$$\Phi(x, y) = e^{\frac{1}{2}\ell og(1-\xi^2) - \frac{\xi}{2\sigma^2(1-\xi^2)}\left[\xi(x-y)^2 - 2(1-\xi)(x-\mu)(y-\mu)\right]}, \tag{36}$$

with $\Psi(x, y)$, the continuous variable version of the π_{ij} of (5) that is given in Theorem 4.4 of Busenberg and Castillo-Chavez (1991), and solve for $\phi(x, y)$. An obvious solution is $\phi(x, y) = \Psi(x, y) = \Phi(x, y)$, which is not very informative, but it is the only one. If we wish to generate an alternative symmetric joint distribution with marginals f(u) such that u is proportional to $N(\mu, \sigma)$, the normal distribution with mean μ and standard deviation σ, we may easily do so by choosing $\phi(x,y)$ with properties reflecting our assumptions about affinity. Say we choose to parameterize ϕ as

$$\phi = \begin{cases} a \geq 0, & \text{if } x = y, \\ b \geq 0, & \text{if } x \neq y, \end{cases} \tag{37}$$

which is a pronounced version of the "like-with-like" functions studied by, e.g., Blythe and Castillo-Chavez (1989), Castillo-Chavez and Blythe (1989), and Castillo-Chavez *et al.* (1991), and known as "diagonal mixing" in the context of sexual mixing models (Blythe *et al.* 1992). Then we have at once that

$$f_{X,Y}(x, y) = f(x)f(y)\Big[K\Big(1-b-(a-b)f(x)\Big)\Big(1-b-(a-b)f(y)\Big) + a\delta(x-y) + b\Big(1-\delta(x-y)\Big)\Big], \tag{38}$$

where

$$K = \frac{2\sqrt{\pi}\sigma}{2\sqrt{\pi}\sigma(1-b)-a}, \tag{39}$$

and $\delta(\cdot)$ is the Dirac delta function. This joint density function is of course quite different from (34), although the marginals are the same. For (38) we find that the correlation coefficient is given simply by

$$\xi = \frac{(a-b)}{2\sqrt{\pi}\sigma}. \tag{40}$$

which may be either positive or negative, depending on whether or not the diagonal parameter, a, dominates. When $a = b = 1$, (37) reduces to $f_{X,Y}(x, y) = f(x)f(y)$, i.e., independent variables, and there is naturally zero correlation between X and Y.

In the above examples we have one or two parameters, varying over specified subranges, describing an infinite family of distributions, all of which have the same marginals. In these cases, these parameters [ν in (31), ξ in (36), and a and b in (38)] act to parameterize the affinity function ϕ (see also Farlie, 1960; Placket, 1965; Johnson and Kotz, 1972; Castillo-Chavez *et al.*, 1991, and the parameterizations that we exploited in Examples 1 and 2 above). We may thus offer the interpretation that the ϕ_{ij} provide a means for describing and implementing the most general form of correlation structure in symmetric bivariate distributions, and that their non-uniqueness with respect to the marginals is no more a restriction than it is with the correlation coefficient ξ in the bivariate

normal case, and subject only to the issues of parameter estimation. We are at present extending the results on symmetric joint distributions to the more general bivariate case where π need not be symmetric, and examining the implications for higher-dimensional problems.

4 Discussion and Conclusions

We have presented a new result in the description of the probability of event pairs where within-pair order is immaterial. Equation (5) sumarizes the underlying key result. Here we suggest four areas where useful applications of (5) may be found. Under the specified conditions for event pair processes--symmetric joined distributions--we may make use of the concept of "affinity" between pairs of events of different types. The elements of the ϕ matrix characterize the degree to which particular types of events appear together in pairs in a nonrandom fashion. The card-game example (Example 1) provides an illustration of how the result, and the ϕ-method, may be applied, but perhaps represents the least interesting area of application. The potential for the method to provide a consistent theoretical basis for the measurement of species association (Example 2) is perhaps one of the most interesting areas of potential application, particularly if other similar applications are exploited. Clearly, work is required both in the general analysis of the association-measure technique, and in the design of experimental protocols to make use of it. Applications in the areas of sexually transmitted disease dynamics and in population genetics (Example 3) are by now appreciated and work is under way to further develop these themes. The final application, Example 4, is particularly interesting, indicating as it does that any symmetric bivariate joint distribution may be represented as a member of the ϕ-family. In particular, this gives us a generalization of the idea of correlation structure, and provides a framework for the construction of bivariate distributions with specified properties when the marginals are known.

The card game also helps illustrate the difference between static and dynamic models for paired events. Assume that the number of red and black cards have their own population dynamics. Let $T_1(t)$ and $T_2(t)$ denote the number of red and black cards respectively, at time t. Changes over time are due to an unspecified stochastic or deterministic model. Let $T(t) \equiv T_1(t)+T_2(t)$ and $\rho_i(t) \equiv T_i(t)/T(t)$, i=1,2. If the deck is fair, then two reds occur with probability $\rho_1(t) \times \rho_1(t)$, two blacks with probability $\rho_2(t) \times \rho_2(t)$, and two mismatched with probability $2 \times \rho_1(t) \times \rho_2(t)$. Consider an unfair deck, where cards have some tendency to remain together. We have two events: E_1 = "red card", and E_2 = "black card", with $\rho_1(t) \equiv T_1(t)/T(t)$ and $\rho_2(t) \equiv T_2(t)/T(t)$. We can then make the following satements:

$$\text{if } \phi = \begin{pmatrix} 1/\rho_1(t) & 0 \\ 0 & 1/\rho_2(t) \end{pmatrix} \quad \text{then} \quad \pi = \begin{pmatrix} \rho_1(t) & 0 \\ 0 & \rho_2(t) \end{pmatrix}, \tag{41}$$

$$\text{if } \phi = \begin{pmatrix} \alpha & \alpha \\ \alpha & \alpha \end{pmatrix},\ \alpha \in [0,1] \text{ then } \quad \pi = \begin{pmatrix} \rho_1^2(t) & \rho_1\rho_2 \\ \rho_1\rho_2 & \rho_1^2(t) \end{pmatrix}. \tag{42}$$

The use of frequency dependent ϕ-values, $\phi_{ij}(t) \equiv [f_i/\rho_i(t)] \times \delta_{ij}$ with constants $f_i \in [0,1]$ and where δ_{ij} denotes the Kronecker delta, reduce Equation (7) to "preferred" mixing, a generalization of Nold's (1980) and Hethcote and Yorke's (1984) mixing (see Jacquez *et al.* 1988):

$$p_{ij} = f_i\delta_{ij} + (1-f_i)\left[\frac{(1-f_j)\rho_j(t)}{\sum_{m \in S}(1-f_m)\rho_m(t)}\right], \tag{43}$$

where i, j =1,2,3,...,n.

In Equation (43) f_i can be interpreted as the reserved fixed (no time dependence) proportion for within type (red or black) suffling while $(1-f_i)$—the non-reserved proportion—which is shuffled at random while a dynamical or stochastic model governs the dynamics of $T_i(t)$, i=1,2. Hence, the selection of $\phi_{ij}(t) \equiv [f_i/\rho_i(t)] \times \delta_{ij}$ gives rise to the above conditional probabilities for event-pairs. Because the ϕ-matrix is symmetric, and condition (9) needs to be satisfied, a game that generates *only* mismatched cards seems only possible (with frequency-dependent ϕ's) when $\phi_{ii} \equiv 0$ and $\phi_{ij} = 1/2$. Hence the situation in which we get only mismatches does not hold if we add dynamics. There are some clear generalizations to multiple types and to event-pairs of event-pairs which correspond, in the parlance of sexually-transmitted diseases (STD's), to heterosexual mixing.

Finally, we must point out that the representation theorem of Busenberg and Castillo-Chavez (1989, 1991) only provides a way of representing an infinite class of possibilities. It is impossible to represent all possible matings with less degrees of freedom than those inherent to the problem (this is true of all representation theorems). The value of a general representation theorem lies on its ability to help understand complex subcases, in this case same-sex pairings. The introduction of "affinities" in terms of our affinity matrix ϕ cannot possible make all paired events scenarios transparent. However, as we show in this paper, these affinities capture familiar patterns of pairing. Furthermore, these affinities provide rich ways of constructing large and flexible families of parametric distributions of paired events. The usefulness of these parametric families cannot be predicted, however, our recent work on "dating" (Blythe 1992, Rubin *et al.* 1992, Luo and Castillo-Chavez 1992, Lubkin *et al.* 1994, Castillo-Chavez *et al.* 1994, Hsu Schmitz and Castillo-Chavez 1994, Hsu Schmitz 1993, Hsu Schmitz *et al.* 1993) shows that the method is indeed promising.

Acknowledgements

S. P. Blythe's research has been partially supported by funds from the Office of the Dean of the College of Agriculture and Life Sciences at Cornell University and the Wellcome Trust. This research

has also been partially supported by NSF grant DMS-9112821 to Stavros Busenberg, and by NSF grant DMS-8906580, DEB-9253570 (Presidential Faculty Fellowship Award) and Hatch Project Grant NYC 151-409, USDA to C. Castillo-Chavez. This research is also partially supported by the U. S. Army Research Office through the Mathematical Sciences Institute at Cornell University (contract DAAL03-91-C-0027). This research was completed while C. Castillo-Chavez was a visiting member of the Isaac Newton Institute, Cambridge University.

References

1. Anderson, R.M. and May, R.M. (1984). Spatial, temporal, and genetic heterogeneity in host populations and the design of immunization progammes. *IMA J. Math. Appl. Med. Biol.* 1, 233-266.
2. Anderson, R.M., Medley, G.F., May, R.M. and Johnson, A.M. (1987). A study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. Appl. Med. Biol.* 3, 229-263.
3. Barbour, A.D. (1978). MacDonald's model and the transmission of bilharzia. *Trans. Roy. Soc. Trop. Med. Hyg.* **72**, 6-15.
4. Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In: *Studies in Item Analysis and Prediction*, Herbert Solomon (ed.) 158-168. Stabford University Press, Stanford, CA.
5. Blythe, S.P. (1992). Heterogeneous sexual mixing in populations with multiple groups. *Mathematical Population Studies.* **3(3)**, 173-188.
6. Blythe, S.P. and Anderson, R.M. (1989). Heterogeneous sexual activity models of HIV transmission in male homosexual populations. *IMA J. Math. Appl. Med. Biol.* **5**, 237-260.
7. Blythe, S.P. and Castillo-Chavez, C. (1989). Like-with-like preference and sexual mixing models. *Math. Biosci.* **96**, 221-238.
8. Blythe, S. P., Castillo-Chavez, C., and Casella, G. (1992) Empirical methods for the estimation of the mixing probabilities for socially-structured populations from a single survey sample. *Mathematical Population Studies.* **3(3)**: 199-225
9. Blythe, S. P., Castillo-Chavez, C., Palmer, J. and Cheng, M. (1991). Towards unified theory of mixing and pair formation. *Math. Biosci.* **107**: 379-405.
10. Busenberg, S. and Castillo-Chavez, C. (1989). Interaction, pair formation and force of infection terms in sexually transmitted diseases. *Lect. Notes Biomath.* 83, 289-300.
11. Busenberg, S. and Castillo-Chavez, C. (1991). A general solution of the problem of mixing of subpopulations, and its application to risk- and age-structured epidemic models. *IMA J. Math. Appl. Med. Biol.* 8, 1-29.
12. Castillo-Chavez, C. and Blythe, S.P. (1989). Mixing framework for social/sexual behavior. *Lec. Notes Biomath.* **83**, 275-288.
13. Castillo-Chavez, C. and Busenberg, S. (1991). On the solution of the two-sex problem. In *Proceedings of the International Conference on Differential Equations and Applications to Biology and Population Dynamics* (S. Busenberg and M. Martelli, eds.). *Lect. Notes Biomath.* **92**, 80-98.

14. Castillo-Chavez, C., Busenberg, S. and Gerow, K. (1991). Pair formation in structure populations. In *Differential Equations With Applications in Biology, Physics and Engineering* (J. Goldstein, F. Kappel and W. Schappacher, eds.). Marcel Dekker, New York, Basel, Hong-Kong, pp. 47-65.

15. Castillo-Chavez, C., Velasco-Hernández, and Fridman S. (1994). Modeling contact structures in biology. In: *Frontiers from Theoretical Biology.* S. A. Levin (ed). Lecture Notes in Biomathematics Vol 100, Springer-Verlag (in the press).

16. Crow, J.F. (1986). *Basic Concepts in Quantitative and Evolutionary Genetics.* W.H. Freeman Co., New York.

17. Dietz, K. and Hadeler, K.P. (1988) Epidemiological models for sexually transmitted diseases. *J. Math. Biol.* **26**, 1-25.

18. Dietz, K. and Schenzle, D. (1985). Proportionate mixing models for age-dependent infection transmission. *J. Math. Biol.* **22**, 117-120.

19. Farlie, D.J.G. (1960). The performance of some correlation coefficients for a general bivariate distribution. *Biometrika* **47**, 307-323.

20. Gumbel, E.J. (1960). Bivariate exponential distributions. *J. Amer. Stat. Assoc.* **55**, 698-707.

21. Hethcote, H.W. and Yorke, J.A. (1984). Gonorrhea transmission dynamics and control. *Lect. Notes Biomath.* **56**,

22. Hubalek, Z. (1973). Coefficients of association and similarity based on arbitrary (presence-absence) data. *Biol. Rev.* **57**, 669-689.

23. Hsu Schmitz S.-F, Busenberg S., and Castillo-Chavez C. (1993). On the evolution of marriage functions: It takes two two tango. *Biometrics Unit Technical Report,* BU-1210-M. Cornell University, Ithaca, New York.

24. Hsu Schmitz S.-F (1993). *Some theories, estimation methods, and applications of marriage functions and two-sex mixing functions in demography and epidemiology.* Ph. D. Dissertation, Cornell University, Ithaca New York.

25. Hsu Schmitz S.-F and Castillo-Chavez C. (1994). Parameter estimation in non-closed social networks related to the dynamics of sexually-transmitted diseases. In: *Modeling the AIDS Epidemic: Planning, Policy, and Prediction.* E. H. Kaplan and M. L. Brandeau (eds.). Raven Press, New York.

26. Jacquez, J.A., Simon, C.P., Koopman, J., Sattenspiel, I. and Perry, T. (1988). Modelling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.* **92**, 119-199.

27. Johnson, N.L. and Kotz, S. (1972). In *Distribution in Statistics, Vol. IV: Continuous Multivariate Distributions.* John Wiley Sons, Inc., New York.

28. Krebs, C.J. (1978). *Ecology: The Experimental Analysis of Distribution and Abundance*, 2nd ed. Harper-Row, New York.

29. Lazarfeld, P. F. (1956). Some observations on dichotomous systems. Sociology Department, Columbia University, New York. Technical report. October 1956.

30. Legendre, L. and Legendre, P. (1983). *Numerical Ecology*. Elsevier, New York.

31. Lubkin S., Hsu Schmitz S.-F, and Castillo-Chavez C. (1994). A framework for modeling inheritance of social traits. In: *Proceedings of the 3rd International Conference on Mathematical Population Dynamics*. O. Arino, D. E. Axelrod, and M. Kimmel (eds.) Wuerz Publishing Ltd. (in the press).

32. Ludwig, J.A. and Reynolds, J.F. (1988). *Statistical Ecology*. Wiley Interscience, New York.

33. Luo X. and Castillo-Chavez, C. (1993). Limit behavior of pair-formation for a large dissolution rate. *Journal of Mathematical Systems, Estimation, and Control* 3:247-264.

33. Mielke, P.W. and Siddiqui, M.M. (1965). A combinatorial test for independence of dichotomous responses. *J. Amer. Stat. Assoc.* **60**, 437-441.

34. Mood, A.M., Graybill, F.A. and Boes, D.C. (1974). *Introduction to the Theory of Statistics*, 3rd ed. McGraw-Hill, New York.

35. Nold, A. (1980). Heterogeneity in disease-transmission modeling. *Math. Biosci.* **52**, 227-240.

36. Plackett, R.L. (1965). A class of bivariate distributions. *J. Amer. Stat. Assoc.* **80**, 516-522.

37. Rubin G., Umbach D., Hsu Schmitz S.-F, and Castillo-Chavez C. (1992). Using mark-recapture methodology to estimate the size of a population at risk for sexaully-transmitted diseases. *Statistics in Medicine* 11:1533-1549.