"On the Construction of Arbitrary m-Dimensional Joint Distributions"

by

S.P. Blythe

Department of Statistics and Modelling Science,
Strathclyde University,
Livingstone Tower,
Glasgow G1 1XH,
Scotland.

and
Carlos Castillo-Chavez

Biometrics Unit,
337 Warren Hall,
Cornell University,
Ithaca, NY 14853
U.S.A.
ABSTRACT

A representation method is given for constructing completely general multidimensional \((m \geq 2)\) joint distributions, given the marginals and a set of arbitrary bounded functions. Any existing continuous, discrete, or mixed type density function can in principle be represented in this manner, but a more useful application is the construction of new forms with specified characteristics, which is relatively straightforward. Proofs and examples are presented.
1. INTRODUCTION

Multivariate distributions other than the multinormal or (a few) other standard forms have an intrinsic interest in distribution theory, as well as some utility in certain applications. For example, in stochastic simulation, and the attendant stochastic analysis, it can be useful to have a specified joint distribution of linked events. While there has been considerable effort in developing bivariate results, particularly in terms of expansions (c.f. Johnson and Kotz, 1972), and some one-parameter "general" joint distributions have been offered (e.g. that of D. Morganstern (Plackett, 1965); see also Farlie, 1960; Gumbel, 1960; Plackett, 1965; Johnson and Kotz, 1972), there has been little practical help in the construction of arbitrary distributions. For dimension greater than two, one must usually fall back on modified-Pearson, or exponential-family, distributions.

Recent theoretical and practical work on the epidemiology of sexually transmitted diseases (greatly fueled by the HIV/AIDS epidemic) has caused considerable attention to be paid to the problem of generating symmetric arbitrary bivariate joint density functions, where the marginals are known (for a review, see Castillo-Chavez, 1989, and Blythe et al. 1990a). In that context the marginal density is the probability density function of sexual activity (new partners per unit time) in a human population, and the conditional density is a "mixing function", describing the probabilities of sexual contact between people of given activities. A number of highly parameterized families have been suggested (op. cit.), one of which (Blythe and Castillo-Chavez, 1989, Castillo-Chavez and Blythe, 1989) was soon extended to give an extremely useful result: a parameterized representation of all viable "mixing functions" was developed by Busenberg and Castillo-Chavez (1989, 1990), Castillo-Chavez and Busenberg (1991). Results for homosexual and heterosexual populations, including age-structure effects, are now available (op. cit.), as well as arbitrarily structured populations (Blythe 1990), and these are just starting to be implemented in practical disease models (c.f. Blythe et al. 1990b).

The existence of the Busenberg/Castillo-Chavez representation of the symmetric bivariate case (activity x activity), and partially symmetric four-dimensional case (activity x activity, age x age), suggested that a further extension of these ideas might have utility outside epidemiology. It has been demonstrated (Blythe et al. 1990) that applications exist in affinity/association analysis, and in statistical distribution theory. This paper enters the latter area, by extending the representation theorem to arbitrary dimensions, and disposing of the requirement that the joint distribution be symmetric. It is shown that any (real) joint distribution may be constructed using this method, given knowledge of the marginals, and a set of arbitrary non-negative functions.
2. MULTIDIMENSIONAL JOINT DENSITY FUNCTIONS

Let \( m \geq 2 \) be the number of dimensions, i.e. the number of independent random variables. Let \( f_i(x_i) \) be the marginal probability density function of the \( i^{th} \) variable, with domain set \( Q_i = \{ x_i : f_i(x_i) > 0 \} \), and let \( Q_i \in \mathbb{R} \) exist. Clearly, \( f_i(x_i) = 0 \) for \( x_i \notin Q_i \). Note that \( x_i \) may thus be of continuous, discrete, or mixed type, depending only upon the continuity of \( Q_i \).

Let \( f_J(x_1,x_2,...,x_m) \) be the joint probability density function of the \( m \) variables, with domain set \( Q_J = \{ x_1,x_2,...,x_m : x_i \in Q_i, \ i \in [1,m] \} \). Again \( f_J(x_1,x_2,...,x_m) = 0 \) for \( (x_1,x_2,...,x_m) \notin Q_J \). A valid joint probability density function must satisfy

\[
f_J(x_1,x_2,...,x_m) \geq 0, \ (x_1,x_2,...,x_m) \in \mathbb{R}^m, \tag{1}
\]

and

\[
\int_{u_1=-\infty}^{+\infty} \int_{u_2=-\infty}^{+\infty} ... \int_{u_{i-1}=-\infty}^{+\infty} \int_{u_{i+1}=-\infty}^{+\infty} \int_{u_m=-\infty}^{+\infty} f_J(u_1,u_2,...,u_{i-1},x_i,u_{i+1},...,u_m) \, du_1 ... du_{i-1} du_{i+1} ... du_m = f_i(x_i), \ (x_1,x_2,...,x_m) \in Q_J, \text{ for all } i \in [1,m]. \tag{2}
\]

Also note that by definition,

\[
\int_{u_1=-\infty}^{+\infty} ... \int_{u_m=-\infty}^{+\infty} f_J(u_1,...,u_m) \, du_1 ... du_m = 1, \tag{3}
\]

and

\[
\int_{u_i=-\infty}^{+\infty} f_i(u_i) \, du_i = 1, \ i \in [1,m]. \tag{4}
\]

We can then construct the following Theorems (see also Busenberg and Castillo-Chavez 1989, 1990, Castillo-Chavez and Busenberg 1990, and Blythe et al., 1990):
**Theorem 1.**

Let $\Psi(x_1, x_2, ..., x_m) : \mathbb{R}^m \rightarrow \mathbb{R}^{2m}$ be an arbitrary function such that $\Psi(x_1, x_2, ..., x_m) > 0$ if $(x_1, x_2, ..., x_m) \in Q_j$, $\Psi(x_1, x_2, ..., x_m) = 0$ if $(x_1, x_2, ..., x_m) \notin Q_j$, and

$$\int_{u_1=-\infty}^{+\infty} \int_{u_2=-\infty}^{+\infty} \cdots \int_{u_{i-1}=-\infty}^{+\infty} \int_{u_{i+1}=-\infty}^{+\infty} \cdots \int_{u_m=-\infty}^{+\infty} \left[ \prod_{j=1}^{m} f_j(u_j) \right] \times
\Psi(u_1, u_2, ..., u_{i-1}, x_i, u_{i+1}, ..., u_m) \, du_1 \cdots du_{i-1} \, du_{i+1} \cdots du_m = 1,$$

for all $i \in [1,m]$. Then

$$f_j(x_1, x_2, ..., x_m) = \Psi(x_1, x_2, ..., x_m) \left[ \prod_{j=1}^{m} f_j(x_j) \right]$$

is a valid joint probability density function.

**Proof.**

We must show that Eq (6) satisfies Eqs (1) – (3). First, note that by the definition of the sets $Q_i$, $i \in [1,m]$, we have

$$\prod_{j=1}^{m} f_j(x_j) \text{ is } \begin{cases} > 0 & \text{if } (x_1, x_2, ..., x_m) \in Q_j \\ = 0 & \text{if } (x_1, x_2, ..., x_m) \notin Q_j \end{cases},$$

and that by definition $\Psi(x_1, x_2, ..., x_m) > 0$ only for $(x_1, x_2, ..., x_m) \in Q_j$. Hence $f_j(x_1, x_2, ..., x_m)$ as given by Eq (6) is positive for $(x_1, x_2, ..., x_m) \in Q_j$, and zero for $(x_1, x_2, ..., x_m) \notin Q_j$, so that Eq (1) is satisfied. Substituting Eq (6) into Eq (2), and extracting the $i^{th}$ marginal, gives
which by Eq (5) is equal to $f_i(x_i)$, so that Eq (2) is satisfied for all $i \in [1,m]$. Using this result, we see that Eq (3) is automatically satisfied by Eq (4). This completes the Proof.

**Theorem 2.**

All valid joint probability density functions may be represented in the form of Eq (6), given $\Psi(x_1,x_2,\ldots,x_m)$ as defined in Theorem 1.

**Proof.**

Let $f_j(x_1,x_2,\ldots,x_m)$ be a valid joint probability density function on $(x_1,x_2,\ldots,x_m) \in Q_J$, with marginal probability density functions $f_i(x_i)$ on $Q_i$, $i \in [1,m]$. Then define

$$
\Psi(x_1,x_2,\ldots,x_m) = \begin{cases} 
\frac{f_j(x_1,x_2,\ldots,x_m)}{\prod_{j=1}^{m} f_j(x_j)}, & (x_1,x_2,\ldots,x_m) \in Q_J \\
0, & (x_1,x_2,\ldots,x_m) \notin Q_J
\end{cases} \quad (8)
$$

We see at once that $\Psi(x_1,x_2,\ldots,x_m) > 0$ for $(x_1,x_2,\ldots,x_m) \in Q_J$, and $\Psi(x_1,x_2,\ldots,x_m) = 0$ for $(x_1,x_2,\ldots,x_m) \notin Q_J$. Substituting Eq (8) into Eq (5), we see that

$$
\frac{1}{f_i(x_i)} \int_{u_1=-\infty}^{+\infty} \int_{u_2=-\infty}^{+\infty} \cdots \int_{u_{i-1}=-\infty}^{+\infty} \int_{u_{i+1}=-\infty}^{+\infty} \int_{u_m=-\infty}^{+\infty} f_j(u_1,u_2,\ldots,u_{i-1},x_i,u_{i+1},\ldots,u_m) \, du_1 \cdots du_{i-1} du_{i+1} \cdots du_m
$$
by Eq (2). Hence $\Psi(x_1, x_2, \ldots, x_m)$ satisfies the constraints defined in Thereom 1, and the Proof is complete.

**Thereom 3.**

The function $\Psi(x_1, x_2, \ldots, x_m)$ may always be represented in the form

$$\Psi(x_1, x_2, \ldots, x_m) = \frac{\prod_{j=1}^{m} R_j(x_j)}{V^{m-1}} + \phi(x_1, x_2, \ldots, x_m),$$

where

$$R_i(x_i) \equiv \int_{u_i=-\infty}^{+\infty} \int_{u_j=-\infty}^{+\infty} \int_{u_{i+1}=-\infty}^{+\infty} \int_{u_m=-\infty}^{+\infty} \left[ \prod_{j=1}^{m} \frac{f_j(u_j)}{f_i(x_i)} \right] x$$

$$\left(1 - \phi(u_1, u_2, \ldots, u_{i-1}, x_i, u_{i+1}, \ldots, u_m)\right) du_1 \ldots du_{i-1} du_{i+1} \ldots du_m,$$

and

$$V \equiv \int_{-\infty}^{+\infty} f_k(u_k) R_k(u_k) \, du_k,$$

for any $k \in [1,m]$, with

$$\phi(x_1, x_2, \ldots, x_m) \geq 0, \quad (x_1, x_2, \ldots, x_m) \in Q_J,$$

and

$$\phi(x_1, x_2, \ldots, x_m) = 0, \quad (x_1, x_2, \ldots, x_m) \notin Q_J.$$
We also have the restrictions $R_i(x_i) \geq 0 \ (i \in [1,m], \ x_i \in Q_j), \ V \geq 0$, and

$$\lim_{R_1 \ldots R_m \to 0} \left[ \prod_{j=1}^{m} R_j(x_j) \right] = 0. \quad (14)$$

**Proof.**

By definition, $\Psi(x_1,x_2,\ldots,x_m) > 0$ for $(x_1,x_2,\ldots,x_m) \in Q_J$, so that it will always be possible to find a set of functions that are less than $\Psi$ at every point in $(x_1,x_2,\ldots,x_m) \in Q_J$. We may generate these functions arbitrarily using $\Theta_i(x_i) \geq 0$ on $(x_1,x_2,\ldots,x_m) \in Q_J$, zero elsewhere, such that

$$\Psi(x_1,x_2,\ldots,x_m) \geq \prod_{j=1}^{m} \Theta_j(x_j). \quad (15)$$

Then, introducing a function $\phi(x_1,x_2,\ldots,x_m) \geq 0 \ ((x_1,x_2,\ldots,x_m) \in Q_J, \ \text{zero elsewhere})$ as the difference between the lower bound of Eq (15) and the value of $\Psi(x_1,x_2,\ldots,x_m)$, we may represent $\Psi(x_1,x_2,\ldots,x_m)$ in the form

$$\Psi(x_1,x_2,\ldots,x_m) = \prod_{j=1}^{m} \Theta_j(x_j) + \phi(x_1,x_2,\ldots,x_m), \quad (16)$$

for all $(x_1,x_2,\ldots,x_m)$. If we now multiply both sides of Eq (16) by $\prod_{j=1}^{m} f_j(x_j)$, and integrate over all $x_j$, $j \neq i$, we are left with $\Theta_i(x_i)$ outside the multiple integral, so that upon rearranging, we see that $\Theta_i(x_i)$ must satisfy

$$\Theta_i(x_i) = \frac{R_i(x_i)}{\prod_{j=1}^{m} \int_{-\infty}^{+\infty} f_j(u_j) \ \Theta_j(x_j) \ du_j}, \ i \in [1,m], \quad (17)$$
where the $R_i(x_i)$ are as defined in Eq (11). A final multiplication by $f_j(x_j)$, and integration over $x_i$, reveals that

$$V = \int_{-\infty}^{+\infty} f_i(u_i) R_i(u_i) \, du_i = \prod_{j=1}^{m} \left[ \int_{-\infty}^{+\infty} f_j(u_j) \Theta_j(x_j) \, du_j \right], \text{ any } i,$$

so that we may write

$$\prod_{i=1}^{m} \left[ \prod_{k=1}^{m} \int_{k \neq i}^{+\infty} f_k(u_k) \Theta_k(x_k) \, du_k \right] = \frac{V^m}{\prod_{i=1}^{m} \left[ \int_{-\infty}^{+\infty} f_i(u_i) \Theta_i(x_i) \, du_i \right]}$$

$$= V^{m-1}.$$ 

Hence we may write the product

$$\prod_{k=1}^{m} \Theta_k(x_k) = \frac{\prod_{k=1}^{m} R_k(x_k)}{V^{m-1}},$$

regaining, through Eq (16), Eq (10) for $\psi(x_1, x_2, \ldots, x_m)$, and the Proof is complete.

**Corollary 1.**

From Theorems 1 - 3 it follows that all multivariate joint probability density functions $f_j(x_1, x_2, \ldots, x_m)$ may be represented in a single generic form, given the $m$ marginal densities $f_i(x_i)$, and arbitrary non-negative functions $\phi(x_1, x_2, \ldots, x_m)$:

$$f_j(x_1, x_2, \ldots, x_m) = \left[ \prod_{j=1}^{m} f_j(x_j) \right] \left[ \prod_{j=1}^{m} \frac{R_j(x_j)}{V^{m-1}} + \phi(x_1, x_2, \ldots, x_m) \right].$$
with \((x_1, x_2, \ldots, x_m) \in Q_f\), and \(R_i(x_j)\) and \(V\) as given in Eqs (11) and (12), respectively.

**Remark 1.**

Because we have defined our formulae in \((x_1, x_2, \ldots, x_m) \in Q_f\), Eq (15) covers continuous, discrete, and mixed distributions. For purely discrete distributions, however, it is usually more convenient to deal directly with sums; the alterations required in the definitions of the \(R_i(x_j)\) and \(V\) are trivial.

**Remark 2.**

Because the representation theorems involve only a lower bound on \(\Psi(x_1, x_2, \ldots, x_m)\), the result Eq (15) should be true whether

\[
\psi^2 = \mathbb{E} \left[ \frac{\psi(x_1, x_2, \ldots, x_m)}{\prod_{j=1}^{m} f_j(x_j)} \right] - 1
\]

is bounded or not.

3. NON-UNIQUENESS OF THE REPRESENTATION.

It is clear that \(\phi(x_1, x_2, \ldots, x_m)\) has more “degrees of freedom” than \(f_j(x_1, x_2, \ldots, x_m)\), being restricted only by inequalities, so that we may expect that there will be many \(\phi(x_1, x_2, \ldots, x_m)\) which will satisfy Eq (15), for given \(f_j(x_1, x_2, \ldots, x_m)\) and \(\{f_j(x_j)\}\). We can demonstrate that this is the case as follows. Say we have some \(\Psi(x_1, x_2, \ldots, x_m)\), and a \(\phi(x_1, x_2, \ldots, x_m)\) which generates it, for given \(\{f_j(x_j)\}\). If we wish to find a \(\phi'(x_1, x_2, \ldots, x_m)\) which generates \(\Psi'(x_1, x_2, \ldots, x_m) = \Psi(x_1, x_2, \ldots, x_m)\), for all \((x_1, x_2, \ldots, x_m) \in Q_f\), we must have

\[
\phi'(x_1, x_2, \ldots, x_m) - \phi(x_1, x_2, \ldots, x_m) = \frac{\prod_{j=1}^{m} R_j(x_j)}{V^{m-1}} - \frac{\prod_{j=1}^{m} R'_j(x_j)}{V'^{m-1}}
\] (16)
for all \((x_1, x_2, \ldots, x_m) \in Q_f\). There is (at least) one family of \(\phi'(x_1, x_2, \ldots, x_m)\), transformations of \(\phi(x_1, x_2, \ldots, x_m)\), such that Eq (16) is satisfied and \(\Psi'(x_1, x_2, \ldots, x_m) = \Psi(x_1, x_2, \ldots, x_m)\):

\[
\phi'(x_1, x_2, \ldots, x_m) = \phi(x_1, x_2, \ldots, x_m) + c^m \prod_{j=1}^{m} R_j(x_j) \frac{1}{V^{m-1}}, 0 \leq c \leq 1,
\]

This may readily be verified by substitution into Eq (16). In Eq (17), the upper limit on \(c\) keeps the \(R_i'(x_j) \geq 0\), as \(R_i'(x_j) = R_i(x_j) \left(1 - c^m\right)\); if \(c = 1\) then \(\phi'(x_1, x_2, \ldots, x_m) = \Psi(x_1, x_2, \ldots, x_m)\), while if \(c = 0\) then \(\phi'(x_1, x_2, \ldots, x_m) = \phi(x_1, x_2, \ldots, x_m)\). Thus every \(\phi(x_1, x_2, \ldots, x_m)\) which generates a specific \(\Psi(x_1, x_2, \ldots, x_m)\) (and hence a specific \(\Psi(x_1, x_2, \ldots, x_m)\)) may be transformed into another, slightly larger, \(\phi(x_1, x_2, \ldots, x_m)\) which generates the same \(\Psi(x_1, x_2, \ldots, x_m)\). It is interesting to consider the outcome of repeating this transformation. If we start with some \(\phi(x_1, x_2, \ldots, x_m)\) and its generated \(\Psi(x_1, x_2, \ldots, x_m)\), say \(\phi(0)\) and \(\Psi\), and denote by \(\phi(k)\) the \(\phi\) obtained by iterating the map Eq (17) \(k\) times, keeping \(c\) constant, then we have

\[
\phi(k+1) - \phi(k) = c^m \prod_{j=1}^{m} R_j^{(k)}(x_j) \frac{1}{V^{(k)m-1}}
\]

\[
= c^m \left(1 - c^m\right)^k \prod_{j=1}^{m} R_j^{(0)}(x_j) \frac{1}{V^{(0)m-1}}
\]

\[
= c^m \left(1 - c^m\right)^k \left(\Psi - \phi(0)\right).
\]

This is a geometric series, so we may write at once that

\[
\phi(k) = \frac{1}{\alpha_k} \Psi + \left(1 - \frac{1}{\alpha_k}\right) \phi(0), \text{ with } \alpha_k = \frac{c^m}{1 - (1 - c^m)^k - c^{2m}}.
\]

or
\[ \phi^{(k+1)} = c_m \Psi + (1 - c_m) \phi^{(k)} , \]

so that \( \phi^{(k)} \to \Psi \) as \( k \to \infty \). Inverting Eq (19), we get a decreasing series

\[ \phi^{(k)} = \gamma \Psi + (1 - \gamma) \phi^{(k+1)} , \quad \gamma = (1 - c_m)^{-1} \]

so that we may establish that \( 0 < \phi'(x_1, x_2, \ldots, x_m) < 1 \), by one step of this inverse recurrence, if

\[ \eta \Psi(x_1, x_2, \ldots, x_m) < \phi(x_1, x_2, \ldots, x_m) < 1 - \eta + \eta \Psi(x_1, x_2, \ldots, x_m) , \]

for some \( 0 < \eta < 1 \), and \( (x_1, x_2, \ldots, x_m) \in Q_f \). Hence, if we start with \( \phi \) outside \([0,1]\), then provided that we can find \( c \) such that the inequality (21) is satisfied, then a single transformation gives us a \( \phi' \) in \([0,1]\). For purposes of parameter estimation (Blythe et al. 1991a) it would be useful to know if a \( 0 \leq \phi' \leq 1 \) always exists.

3.1 Independent Joint Density

A case of interest is the independent form of joint probability density function,

\[ f_j(x_1, x_2, \ldots, x_m) = \prod_{j=1}^{m} f_j(x_j) . \]

**Theorem 4.**

For \( 0 \leq \phi(x_1, x_2, \ldots, x_m) \leq 1 \), all joint probability density functions of the form Eq (22) arise from \( \phi(x_1, x_2, \ldots, x_m) \) which are given by

\[ \phi(x_1, x_2, \ldots, x_m) = 1 - \prod_{j=1}^{m} \omega_j(x_j) , \]
where \( 0 \leq \omega_i(x_i) \leq 1 \), \( i \in [1,m] \), and \((x_1,x_2,\ldots,x_m) \in Q_J\).

**Proof.**

First, define the non-negative function

\[
\omega(x_1,x_2,\ldots,x_m) \equiv 1 - \phi(x_1,x_2,\ldots,x_m), \quad (x_1,x_2,\ldots,x_m) \in Q_J, \tag{24}
\]

for \( 0 \leq \phi(x_1,x_2,\ldots,x_m) \leq 1 \), and

\[
\Omega_i(x_i) \equiv \int_{u_1=-\infty}^{+\infty} \int_{u_2=-\infty}^{+\infty} \cdots \int_{u_{i-1}=-\infty}^{+\infty} \int_{u_{i+1}=-\infty}^{+\infty} \int_{u_m=-\infty}^{+\infty} \left[ \prod_{j=1}^{m} f_j(u_j) \right] \times
\omega(u_1,u_2,\ldots,u_{i-1},x_i,u_{i+1},\ldots,u_m) \, du_1 \cdots du_{i-1} du_{i+1} \cdots du_m, \tag{25}
\]

so that

\[
R_i(x_i) = \Omega_i(x_i), \tag{26}
\]

and

\[
V = \int_{-\infty}^{+\infty} f_i(u_i) \Omega_i(u_i) \, du_i, \text{ any } i. \tag{27}
\]

Then, noting that Eq (22) requires

\[
\frac{f_j(x_1,x_2,\ldots,x_m)}{\prod_{j=1}^{m} f_j(x_j)} = 1, \tag{28}
\]

we substitute Eq (26) – (28) into Eq (15), to see that Eq (28) requires that the \( \omega(x_1,x_2,\ldots,x_m) \) satisfy
Clearly Eq (29) implies that \( \omega(x_1, x_2, \ldots, x_m) \) is separable, i.e.

\[
\omega(x_1, x_2, \ldots, x_m) \equiv \prod_{j=1}^{m} \omega_j(x_j),
\]

so that Eq (24) follows, and the Proof is complete (see also Blythe et al., 1990).

4. EXAMPLES.

4.1 Bivariate Distributions.

Of particular interest is the bivariate case \((m = 2)\) of Eq (15). Here the marginals are \( f_X(x) \) and \( f_Y(y) \), and the joint density function is \( f_{X,Y}(x,y) \), given by

\[
f_{X,Y}(x,y) = f_X(x) f_Y(y) \left[ \frac{R_X(x) R_Y(y)}{+\infty} + \phi(x,y) \right] \int_{-\infty}^{+\infty} f_X(u) R_X(u) \, du,
\]

where

\[
R_X(x) = \int_{-\infty}^{+\infty} f_Y(v) \left(1 - \phi(x,v)\right) \, dv,
\]

and
\[ R_Y(y) = \int_{-\infty}^{+\infty} f_Y(u) \left( 1 - \phi(u, y) \right) du \]  

(note that the denominator in Eq (31) could equally well be in terms of the Y-variables). This result is implicit in that of Busenberg and Castillo-Chavez (1990) for “two-sex mixing”.

**Example 1.**

A useful particular case (Blythe and Castillo-Chavez 1990b, Blythe et al. 1990) is the so-called “diagonal” result when \( \phi(x, y) \) is given by

\[ \phi(x, y) = \delta(x-y) a + \left( 1 - \delta(x-y) \right) b , \]

where \( 0 \leq a, b \leq 1 \), and \( \delta(z) \) is a Dirac delta function. Then Eq (31) is simply

\[ f_{j(x,y)} = f_X(x) f_Y(y) \left[ \frac{\left\{ 1 - b - (a-b) f_X(x) \right\}\left\{ 1 - b - (a-b) f_Y(y) \right\}}{1 - b - (a-b) S} + \delta(x-y) a + \left( 1 - \delta(x-y) \right) b \right] , \]

where

\[ S = \int_{-\infty}^{+\infty} f_X(u) f_Y(u) du . \]

(Note that \( x \) and \( y \) appear, in Eq (35), as the argument for the other marginal density function. Depending on the domain of each marginal, this is how mixed distributions can arise even when the marginals are continuous). The use of delta functions with continuous marginals means that we must be careful in the choice of \( a \) and \( b \), as \( R(\cdot) \) must be non-negative; from Eq (35), we see that this implies

\[ \frac{1 - b}{a - b} \geq \text{MAX}\left\{ \text{SUP}\{f_X(x)\}, \text{SUP}\{f_Y(y)\} \right\} , \]
so that we require $a \geq b$ and $b \neq 1$. For example, if both $f_X(x)$ and $f_Y(y)$ are given by $f(x)$, a normal distribution with mean $\mu$ and variance $\sigma^2$, Eq (37) implies that $a$, $b$ and $\sigma^2$ must satisfy

$$\sigma^2 \geq \frac{1}{2\pi} \left( \frac{a-b}{1-b} \right)^2, \quad 0 < b < a < 1. \quad (38)$$

In this case the product-moment correlation coefficient of the joint density function Eq (35) may be shown to be (Blythe et al. 1990)

$$\rho = \frac{(a-b)}{\sqrt{4\pi\sigma^2}}, \quad (39)$$

with range (from Eq (38))

$$0 \leq \rho \leq \frac{(1-b)}{\sqrt{2}} < 0.707107. \quad (40)$$

The regression of $Y$ on $X$ in this example is given by

$$\hat{y} = \mathbb{E}[Y \mid X] = 1 - b - (a-b)\mu + (a-b) (x-1) f(x)$$

$$= \left(1 + \rho\sqrt{4\pi\sigma^2} + a(\mu-1)\right) + \sqrt{2} \rho (x-1) \exp \left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad (41)$$

where the same restrictions on $a$, $b$, $\rho$ and $\sigma^2$. Eq (41) is highly nonlinear, possessing two turning points:

$$x = \frac{1}{2} \left[ \mu + 1 \pm \sqrt{(\mu-1)^2 + 4\sigma^2} \right]. \quad (42)$$

The symmetric bivariate case ($f_j(x,y) = f_j(y,x)$, $f_X(x)$ and $f_Y(y)$ have the same distribution) has been much studied in recent years in the epidemiological literature, where the conditional density function $f_j(x,y)$ is known as a "mixing function", describing the distribution of rates of sexual contact within a population. A number of new specific parametric forms for this conditional have been found,
some of which may be of use in the context of constructing symmetric joint distributions (see Blythe and Castillo-Chavez (1990a), Blythe (1991), and Blythe et al. (1991b) for the functions $\phi$ which generate all of these mixing functions). Blythe and Castillo-Chavez (1989), Castillo-Chavez et al. (1989), and Castillo-Chavez and Blythe (1989) have looked at the symmetric case of Eqs (31)–(33) (i.e. when the two marginals are the same function, and $\phi(y,x) = \phi(x,y)$, so that $R_X(x)$ and $R_Y(y)$ are the same function), for marginal $f(x)$ given by a first-order Erlang distribution, and a variety of $\phi(x,y)$ of the form $\phi(x - y)$, including double exponential, normal, and log-normal forms.

### 4.2 Multivariate Case.

**Example 2.**

For general $m$, an example that is easy to calculate arises when the marginals are all functions on the unit interval (e.g. Beta distributions), and a sufficient number of moments exist. Then a convenient choice of $\phi(x_1, x_2, \ldots, x_m)$ is

$$
\phi(x_1, x_2, \ldots, x_m) = \prod_{j=1}^{m} x_j^{n_j},
$$

(43)

where the $n_j$ are non-negative integers. As all $x_j$ lie in $(0,1)$, so do the $\phi(x_1, x_2, \ldots, x_m)$. Writing

$$
\kappa_j = \mathbb{E}[X_j^{n_j}] = \int_0^1 u_j^{n_j} f_j(u_j) \, du_j, \quad j \in [1, m],
$$

(44)

the $n_j$th moment of the $j$th marginal, we have

$$
f_j(x_1, x_2, \ldots, x_m) = \left( \prod_{j=1}^{m} f_j(x_j) \right) \left[ \prod_{j=1}^{m} \left\{ \frac{1 - (1 - V)^x_j}{\kappa_j} \right\} x_j^{n_j} + \prod_{j=1}^{m} x_j^{n_j} \right].
$$

(45)

where

$$
V = 1 - \prod_{k=1}^{m} \kappa_k.
$$

(46)
Note that if we use $\phi(x_1, x_2, \ldots, x_m)$ given by one minus the RHS of Eq (43), then by Eq (23) the $f_j(x_1, x_2, \ldots, x_m)$ would consist only of the product of the marginals.

The matrix of correlations between pairs of variables, given Eq (45), is then simply

$$
\rho_{i,j} = \begin{cases} 
1 & , i = j \\
\frac{(1 - V)}{\sqrt{\sigma_i \sigma_j}} \left( \mu_i - \frac{\kappa_i^{+1}}{\kappa_i} \right) \left( \mu_j - \frac{\kappa_j^{+1}}{\kappa_j} \right) & , i \neq j 
\end{cases}
$$

(47)

where

$$
\mu_j = \mathbb{E}[X_j] ,
$$

(48)

and

$$
\sigma_j^2 = \mathbb{E}[X_j^2] - \mu_j^2 ,
$$

(49)

are simply the means and variances of the variables, and

$$
\kappa_j^{+1} = \mathbb{E}[X_j^\kappa_j^{+1}] .
$$

(50)

Thus, at least the first, second, $n_j$th, and $n_{j+1}$th moments for each variable $x_j$ must exist. As a trivial example let all the marginals in Eq (45) be uniform on $(0,1)$. Then Eq (47) becomes

$$
\rho_{i,j} = \begin{cases} 
1 & , i = j \\
\frac{12 \ n_i \ n_j}{(2 + n_j) (2 + n_j) \left\{ \prod_{k=1}^{m} (1 + n_k) - 1 \right\}} & , i \neq j 
\end{cases}
$$

(51)

so that correlation is always positive. If all the $n_j$ some $n$, then the $\rho_{i,j}$ (i $\neq$ j) are all equal and positive, but tend rapidly towards zero as $m$ (the number of variables) increases, or if $n$ becomes large.
5. CONCLUSIONS

The method presented appears to be completely general, in that the representation theorem (Corollary 1) states that any joint density function, of any type, may be expressed in the form of Eq (15). To construct a joint function of any desired form, then, one must know the m marginals, and choose a ϕ(x₁,x₂,...,xₘ) which produces the required characteristics (e.g. correlation and variance-covariance matrices, number of free parameters, regression characteristics). It is considerably more difficult to find the ϕ(x₁,x₂,...,xₘ) (other than the trivial solution ϕ(x₁,x₂,...,xₘ) = ψ(x₁,x₂,...,xₘ)) which exactly reproduce a known equation for fₙ(x₁,x₂,...,xₘ), even when the marginals are known (for some known examples, see Blythe 1990, Blythe and Castillo-Chavez 1990a, Blythe et al. 1990), but is in principle possible for all distributions. However, unless an alternative ϕ-representation to ϕ = ψ has some advantage, such as more readily interpretable or estimable parameters, there seems little reason to construct one. The construction, ab initio, of new multivariate forms seems to be a much more useful application of the representation theorem, Eq (15).

The problem of estimation of the ϕ(x₁,x₂,...,xₘ) has yet to be addressed in great detail; some preliminary heuristic results, in the context of sexual mixing, have been presented (Blythe et al. 1991a), and the considerably easier problem of estimation in the context of ecological associations or affinities has begun to be addressed (Blythe et al. 1990).

Finally, there is the question of how best to generate random variates from a joint distribution as specified by Eq (15). This is an open problem, which is the subject of further investigation.
ACKNOWLEDGEMENTS

This research has been partially supported by NSF grant DMS-8906580, NIAID Grant R01 A129178-01, and Hatch project grant NYC 151-409, USDA (to Carlos Castillo-Chavez), by the Office the Dean of the College of Agriculture and Life Sciences at Cornell University, and by the Wellcome Trust. I would like to thank Carlos Castillo-Chavez for the many valuable comments he is about to make on this paper.
REFERENCES


