Statistical Science
2001, Vol. 16, No. 3, 249–274

Conditionally Specified Distributions: An Introduction

Barry C. Arnold, Enrique Castillo and José María Sarabia

Abstract. A bivariate distribution can sometimes be characterized completely by properties of its conditional distributions. The present article surveys available research in this area. Questions of compatibility of conditional specifications are addressed as are characterizations of distributions based on their having conditional distributions that are members of prescribed parametric families of distributions. The topics of compatibility and near compatibility of conditional distributions are discussed. Estimation strategies for conditionally specified distributions are summarized. Additionally, certain conditionally specified densities are shown to provide convenient flexible conjugate prior families in certain multiparameter Bayesian settings.

Key words and phrases: Compatibility, near-compatibility, normal conditionals, conjugate priors, functional equations, exponential families.

CONTENTS

1. A Role for Conditional Specification
2. Compatible Conditional Densities
3. Conditionals in Prescribed Families
4. Near Compatibility
5. Anil Bhattacharyya's Distribution
6. Conditionals in Exponential Families
7. Multivariate Extensions
8. Estimation
9. A Bayesian Niche
10. The Problem of Marginal and Conditional Specification
11. Envoi

1. A ROLE FOR CONDITIONAL SPECIFICATION

A bivariate density is arguably only really understandable in terms of its conditional densities (i.e., normalized cross sections). Efforts to model two-dimensional populations will probably be most easily implemented if they involve modelling assumptions about conditional densities as contrasted with assumptions about marginal densities. For example, if we wish to model the joint distribution of heights and weights of women in a particular human population, we may find it quite plausible to contemplate a unimodal distribution of $X =$ height for a given $Y = \log(\text{weight})$, undoubtedly with the modal value being an increasing function of log(weight). Likewise it is reasonable to expect a unimodal distribution of log(weights) for a given height. Even if we assume that these conditional densities are all normal, we are still quite distant from an assumption of classical bivariate normality with its familiar elliptical contours. Indeed we might even be a little taken aback by the mathematical consequences of assuming the classical bivariate normal model. For example, it guarantees that (height) + log(weight) will be normally distributed. Is that self-evident? Note that unimodal conditionals, of X given Y and of Y given X, certainly do not guarantee unimodal distributions for $X + Y$ and $X - Y$! Even if we assume normal conditionals, we may still encounter a bimodal distribution for $X + Y$ or $X - Y$ (more on this in Section 5).

If we try to do our modelling in terms of marginal densities, we will undoubtedly encounter difficulty. The following exercise (borrowed from Arnold, Castillo and Sarabia, 1999) is designed to underline the difficulties we will likely have in visualizing

Barry C. Arnold is Professor, Department of Statistics, University of California, Riverside, California 92521-0002 (e-mail: barry.arnold@ucr.edu). Enrique Castillo is Professor, Department of Applied Mathematics and Computational Sciences, Universities of Cantabria and Castilla-La Mancha, Spain. José María Sarabia is Professor, Department of Economics, University of Cantabria, Spain.

marginal densities even for known joint densities. A topographical map of a country, showing height above sea level can be viewed, after normalization, as a bivariate density. We claim that the marginal densities in such settings are basically uninformative. In contrast, a few conditional densities (cross sections) can be effectively used to visualize the joint density. Specifically consider three countries with whose topography you are reasonably familiar: perhaps Mexico, Spain and United States. You could identify the three countries quite easily from the joint densities (i.e., the topographic maps). However, the marginal densities are next to useless as tools for identifying the countries (i.e., the joint densities). Figure 1 provides the Y marginals for the three countries in random order. Figure 2 provides the X marginals in the same random order as in Figure 1. Can you sort them out? Figure 10, later in the article, provides the corresponding topographic maps in the same order as in Figures 1 and 2 so that you can grade your ability to visualize marginal densities. In contrast, a few cross sections would easily permit discrimination between the maps of the three countries.

Some role for conditional specification of joint densities thus seems justified. There are several questions that arise in this context. How can we be sure that given families of conditional densities are compatible? In other words, are we sure that there exists any density with conditional densities equal to the given ones? If existence is verified, what about uniqueness? Section 2 will address these questions. We will then turn in Section 3 to situations where the conditional densities are not completely prescribed but are posited to be members of given parametric families of densities. The prototypical example of this, first studied by Bhattacharyya (1943) (shown in Figure 3), involves distributions for (X, Y) such that all conditionals of X given $Y = y$ and of Y given $X = x$ are normal. They form an interesting extension of the family of classical bivariate normal densities and will be discussed at some length in Section 5. Section 4 will address the issue of near-compatibility, focussing for simplicity on the finite discrete case. The family of normal distributions is the quintessential exponential family. It is natural, and straightforward to seek versions of Bhattacharyya's model that involve conditionals in arbitrary exponential families. This will be described in Section 6. Section 7 outlines briefly certain extensions to higher dimensions.

One of the disquieting features of many conditionally specified models is the presence of an awkward normalizing constant which usually must be evaluated numerically. This complicates the business

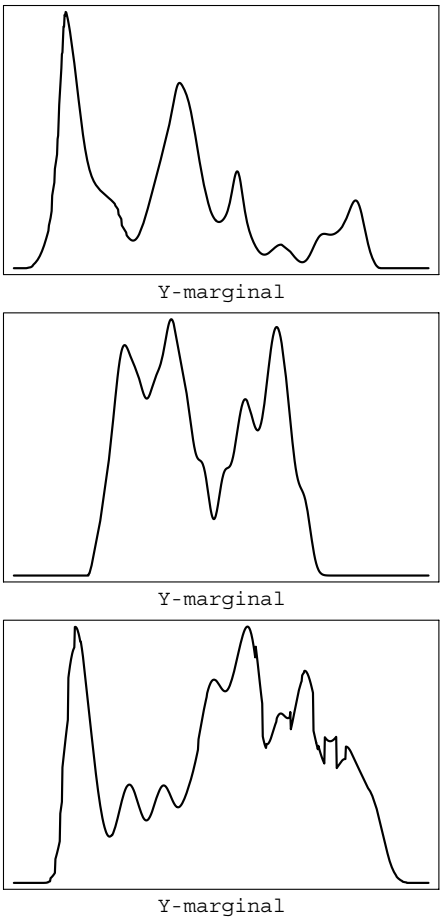

FIG. 1. *Y-marginals (from north to south coast) for Mexico, Spain and the United States in random order.*

of estimating parameters, but several approaches described in Section 8 can be successfully implemented. Conditionally specified distributions turn out to provide convenient conjugate prior families in some multiparameter Bayesian settings. This perhaps unexpected home for such distributions is explained by the fact that conditionally specified densities are "tailor-made" for Gibbs sampling simulation algorithms. A brief discussion of this topic is the subject of Section 9. The final section of the paper provides a discussion of the general problem of marginal and conditional specification in higher dimensions.

The present paper will supply only a survey of ideas relating to "conditional specification of statistical models." If it piques your attention, a good starting point for a more leisurely and detailed overview of the area together with an extensive bibliography is to be found in Arnold, Castillo and Sarabia (1999).

Kotz, Balakrishnan and Johnson (2000) also include discussion of several conditionally specified

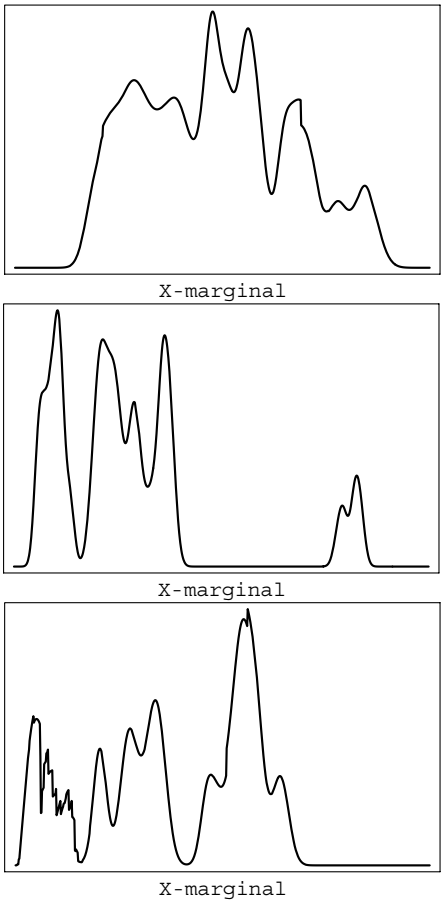

FIG. 2. *X-marginals (from west to east coast) for Mexico, Spain and the United States in random order.*

models in their survey of continuous multivariate distributions.

2. COMPATIBLE CONDITIONAL DENSITIES

Assume that (X, Y) is a random vector that has a joint density with respect to some product measure $\mu_1 \times \mu_2$ on $S(X) \times S(Y)$, where $S(X)$ denotes the set of possible values of X and $S(Y)$ the set of possible values of Y. For example, one variable could be discrete and the other absolutely continuous with respect to Lebesgue measure. The marginal, conditional and joint densities are denoted by $f_X(x), f_Y(y), f_{X|Y}(x|y), f_{Y|X}(y|x), f_{X,Y}(x, y)$ and the sets of possible values $S(X)$ and $S(Y)$ can be finite, countable or uncountable.

Consider two candidate families of conditional densities $a(x, y)$ and $b(x, y)$. We ask when is it true that there will exist a joint density for (X, Y) such that

$$f_{X|Y}(x|y) = a(x, y), \qquad x \in S(X),\ y \in S(Y)$$

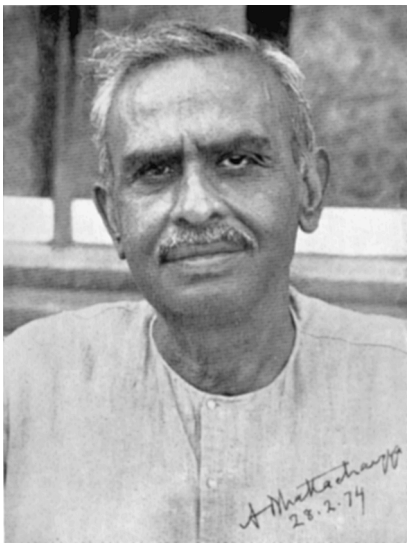

FIG. 3. *A. Bhattacharyya.*

and

$$f_{Y|X}(y|x) = b(x, y), \qquad x \in S(X),\ y \in S(Y).$$

If such a density exists we will say that a and b are compatible families of conditional densities. We define

$$N_a = \{(x, y): a(x, y) > 0\}$$

and

$$N_b = \{(x, y): b(x, y) > 0\}.$$

The following compatibility theorem may be found in Arnold and Press (1989), though it is quite possibly much older than this.

THEOREM 1 (Compatible conditionals). *A joint density $f(x, y)$, with $a(x, y)$ and $b(x, y)$ as its conditional densities, will exist iff*

(i)

$$N_a = N_b = N, \quad \text{say}$$

and

(ii) *There exist functions $u(x)$ and $v(y)$ such that for every $(x, y) \in N$ we have*

$$\frac{a(x, y)}{b(x, y)} = u(x)v(y) \tag{1}$$

in which $u(x)$ is integrable, that is,

$$\int_{S(X)} u(x)d\mu_1(x) < \infty.$$

PROOF. The result is almost self-evident. If $a(x, y)$ and $b(x, y)$ are compatible, then appropriate marginal densities $f(x)$ and $g(y)$ must exist. Then, (1) must hold with $f(x) \propto u(x)$ and

$g(y) \propto [v(y)]^{-1}$. The integrability condition is needed to guarantee that the joint density is proper (integrable). □

EXAMPLE 1 (A compatible case). Consider $S(X) = S(Y) = (0, \infty)$ and

$$a(x, y) = (y + 2)e^{-(y+2)x} I(x > 0),$$

$$b(x, y) = (x + 3)e^{-(x+3)y} I(y > 0).$$

The ratio $a(x, y)/b(x, y)$ is readily confirmed to factor in the form (1) *with* $u(x) = (x + 3)^{-1}e^{-2x}$ $I(x > 0)$*, which is indeed integrable.*

EXAMPLE 2 (Compatible with an improper joint density). Consider $S(X) = S(Y) = (0, \infty)$ and

$$a(x, y) = 2xy^2 I(0 < x < y^{-1}) I(y > 0),$$

$$b(x, y) = 2yx^2 I(0 < y < x^{-1}) I(x > 0).$$

Here $N_a = N_b = N = \{(x, y)\colon x > 0,\ y > 0,\ xy < 1\}$ and the ratio

$$\frac{a(x, y)}{b(x, y)} = \frac{y}{x}$$

factors readily into a function of x times a function of y. However, in this example $u(x) = 1/x$ is not integrable over $(0, \infty)$. So, there is no proper joint density with the given family of conditional densities.

Incompatible examples, those in which $a(x, y)/b(x, y)$ does not factor, are of course very easy to write down. Once we have determined that $a(x, y)$ and $b(x, y)$ are compatible we need to address the question of whether there is a unique joint density compatible with them. An early reference on this problem is Gourieroux and Montfort (1979). The problem may be viewed as one relating to Markov chains. The idea (familiar to Gibbs sampling afficionados) involves starting with X, then generating a value for Y using $b(X, y)$, then generating another X value using $a(x, Y)$, etc.

If $a(x, y)$ and $b(x, y)$ are compatible there will exist a compatible marginal density for X, say $\tau(x)$. We will relate this density $\tau(x)$ to the Markov chain implicitly described above which involves using $b(x, y)$ then $a(x, y)$ to obtain a new X value from an initial X value. The corresponding stochastic kernel, denoted by $ba(\cdot|\cdot)$, is

$$(2) \qquad ba(x|z) = \int_{S(Y)} a(x, y)b(z, y)\, d\mu_2(y).$$

Consider a Markov chain with state space $S(X)$ and transition kernel ba as in (2).

It is evident that a and b will be compatible iff $\tau(x)$ (the compatible marginal for X) is a stationary distribution for this chain. The stationary distribution (long run distribution) will be unique provided that the Markov chain is indecomposable.

EXAMPLE 3 [$a(x, y)$ and $b(x, y)$ compatible but with a nonunique compatible density]. Define the sets

$$A_1 = \{(x, y)\colon -1 < x < 0, -1 < y < 0\}$$

and

$$A_2 = \{(x, y)\colon 0 < x < 1, 0 < y < 1\}$$

and set

$$(3) \qquad a(x, y) = b(x, y) = I((x, y) \in A_1 \cup A_2).$$

It is not difficult to verify that any joint density of the form

$$f(x, y) = \frac{\lambda}{2} I((x, y) \in A_1) + \frac{(1 - \lambda)}{2} I((x, y) \in A_2)$$

for $\lambda \in (0, 1)$, will be compatible with (3). It is also not difficult to verify that the Markov chain with transition kernel ba defined using (3) is decomposable. The state space $S(X) = (-1, 0) \cup (0, 1)$ is a disjoint union of two closed subsets of states, namely $(-1, 0)$ and $(0, 1)$.

The simplest sufficient condition for indecomposibility of the Markov chain with kernel ba is a "positivity" condition. The assumption that $N_a = N_b = S(X) \times S(Y)$ will suffice since it assures us that the kernel ba in (2) will be positive for every x and every z.

EXAMPLE 4 (An alternative approach to conditional specification). It was observed by Castillo and Galambos (1987) that the function

$$(4) \quad F(x, y) = 1 - \exp(-x^\gamma y^\eta), \quad x > 0,\ y > 0$$

is remarkable in that for each fixed $y > 0$, it is a valid distribution function when considered as a function of x. In addition, for each fixed $x > 0$, (4) is a valid distribution function as a function of y. A possible conditional distribution specification is evident here.

We may ask whether it is possible to have a joint distribution for (X, Y) that will have both the families of conditional distributions of X given $Y = y$ and of Y given $X = x$, given by (4). If this is to be true, we will be able to get the corresponding conditional densities of X given Y and of Y given X by

differentiation. For compatibility (using Theorem 1) we would need to consider

$$\frac{f_{X|Y}(x|y)}{f_{Y|X}(y|x)} = \frac{y^{\eta}\gamma x^{\gamma-1}\exp(-x^{\gamma}y^{\eta})}{x^{\gamma}\eta y^{\eta-1}\exp(-x^{\gamma}y^{\eta})} = \frac{\gamma}{\eta}\frac{y}{x} \tag{5}$$

for $x > 0$, $y > 0$.

Evidently (5) may be factored into $u(x)v(y)$. Unfortunately, the function $u(x) = x^{-1}I(x > 0)$ is not integrable, so just as in Example 2, no proper joint distribution exists with conditional distributions given by (4). It is interesting to speculate whether it is ever possible to have $P(X \le x|Y = y) = P(Y \le y|X = x)$ for every $x > 0, y > 0$, as was postulated abortively in this example.

3. CONDITIONALS IN PRESCRIBED FAMILIES

Consider a k-parameter family of densities in $\mathbb{R}$ with respect to μ_1 denoted by $\{f_1(x;\underline{\theta})\colon \underline{\theta} \in \Theta\}$, where $\Theta \in \mathbb{R}^k$. Also consider a second l-parameter family of densities $\{f_2(y;\underline{\tau})\colon \underline{\tau} \in T\}$, where $T \subset \mathbb{R}^l$. In some cases, f_1 and f_2 will be the same but generally they can be different. We wish to identify, if possible, all of the joint densities for a random variable (X, Y) which have all their conditional densities given by f_1 and f_2. Thus, we insist that for every $y \in S(y)$ we have

$$f_{X|Y}(x|y) = f_1(x;\underline{\theta}(y)) \tag{6}$$

and for every $x \in S(X)$ we have

$$f_{Y|X}(y|x) = f_2(y;\underline{\tau}(x)) \tag{7}$$

for certain functions $\underline{\theta}\colon S(Y) \to \Theta$ and $\underline{\tau}\colon S(X) \to T$.

If (6) and (7) are to hold, there must exist marginal densities for X and Y (say, $f_X(x)$ and $f_Y(y)$) such that, writing the joint density as a product of a marginal and a conditional density in both possible ways,

$$f_Y(y)f_1(x;\underline{\theta}(y)) = f_X(x)f_2(y;\underline{\tau}(x)) \qquad \forall x \in S(X),\ y \in S(Y). \tag{8}$$

To solve our problem we must solve the functional equation (8) for $\underline{\theta}(y)$, $\underline{\tau}(x)$, $f_X(x)$ and $f_Y(y)$, for given choices of f_1 and f_2.

Functional equations are notorious for being easy to write down and, usually, being very difficult to solve. In Sections 5 and 6, we will consider certain cases in which such a functional equation can be solved. Before that, we digress to consider the question of near compatibility as opposed to exact compatibility.

4. NEAR COMPATIBILITY

The concepts of compatibility and near compatibility of conditional densities are most readily discussed in the case where X and Y each have only a finite number of possible values. We will restrict discussion to that case. Some extensions to more general settings are possible but more difficult programming problems will be encountered. In the finite case, matrix theory, linear programming and some elementary iterative algorithms will carry us a long way.

We focus then on random variables X and Y with possible values $x_1, x_2, \ldots, x_I$ and $y_1, y_2, \ldots, y_J$, respectively. A possible conditional model for the joint distributions of (X, Y) will consists of two $I \times J$ matrices A and B. Matrix A will have columns which sum to 1, while matrix B has rows summing to 1.

We will say that A and B are compatible if there exists an $I \times J$ matrix P with nonnegative elements, p_{ij}, such that

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1,$$

and for every i and j,

$$a_{ij} = \frac{p_{ij}}{p_{\cdot j}} \tag{9}$$

and

$$b_{ij} = \frac{p_{ij}}{p_{i\cdot}}, \tag{10}$$

where $p_{i\cdot} = \sum_{j=1}^{J} p_{ij}$ and $p_{\cdot j} = \sum_{i=1}^{I} p_{ij}$. If such a matrix P exists, then we can identify, for some random vector (X, Y),

$$p_{ij} = P(X = x_i, Y = y_j),$$
$$a_{ij} = P(X = x_i|Y = y_j)$$

and

$$b_{ij} = P(Y = y_j|X = x_i)$$

for $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$.

Paralelling the discussion in Section 2 we may define $N_A = \{(i, j)\colon a_{ij} > 0\}$ and $N_B = \{(i, j)\colon b_{ij} > 0\}$. Then A and B are compatible iff $N_A = N_B$ and there exist vectors $\underline{u}$ and $\underline{v}$ for which

$$\frac{a_{ij}}{b_{ij}} = u_i v_j \qquad \forall (i, j) \in N_A. \tag{11}$$

If (11) holds, then $\underline{u}$ normalized to sum to 1, will be a compatible marginal density for X.

One approach to determining whether vectors $\underline{u}$ and $\underline{v}$ exist to satisfy (11), involves two closely

related Markov chains. Consider an I state Markov chain with transition matrix BA' and J state Markov chain with transition matrix $A'B$. Denote the long-run distributions of the chains corresponding to BA' and $A'B$ by $\underline{\pi}$ and $\underline{\eta}$, respectively. These long-run distributions are obtainable by solving two systems of linear equations. If A and B are compatible, then

$$a_{ij}\eta_j = b_{ij}\pi_i, \qquad i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, J. \tag{12}$$

Incompatibility of A and B will be signaled by some inequalities in (12). So, one way to resolve the compatibility issue is to obtain the long-run distributions $\underline{\pi}$ and $\underline{\eta}$ and check to see if (12) holds. There is good news. The two systems of equations that we must solve to get the long-run distributions are

$$\underline{\pi}BA' = \underline{\pi} \tag{13}$$

and

$$\underline{\eta}A'B = \underline{\eta}. \tag{14}$$

In fact only one system needs to be solved because the solutions are related by

$$\underline{\eta} = \underline{\pi}B. \tag{15}$$

We remark that solutions to (13) and (14) will always exist [and satisfy (15)], but it is only in the case of compatibility that (12) holds.

This has a Gibbs sampler interpretation. If we try to simulate the joint distribution of (X, Y) using an incompatible pair of conditional probability distribution matrices A and B, we will get different joint distributions depending on whether we start with an X or a Y value.

Compatibility must therefore be checked before we crank up our Gibbs algorithms.

Another view of the compatibility issue is as follows. In seeking a matrix $P = (p_{ij})$ to satisfy (9) and (10) we are seeking solutions to a system of linear equations in the p_{ij}'s, namely,

$$p_{ij} = a_{ij}\sum_{i=1}^{I} p_{ij} \qquad \forall i, j \tag{16}$$

and

$$p_{ij} = b_{ij}\sum_{j=1}^{J} p_{ij} \qquad \forall i, j. \tag{17}$$

These equation must be solved subject to the following constraints:

$$p_{ij} \geq 0 \qquad \forall i, j \tag{18}$$

and

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1. \tag{19}$$

It is possible to identify all the possible solutions to a system of linear constraints like (16), (18) and (19) subject to the nonnegativity requirement (18) using results in Castillo, Cobo, Jubete and Pruneda (1999). Details for the present case may be found in Arnold, Castillo and Sarabia (1999, pages 26–30). However, if we will be happy finding one solution or perhaps a "near" solution in an incompatible case, the following approach reduces our problem to a standard linear programming exercise.

We will introduce a concept called ε-compatibility. We say that A and B are ε-compatible if the following system of equations and inequalities has a solution for $\varepsilon' \geq \varepsilon$ but does not have one for $\varepsilon' < \varepsilon$:

$$\left| p_{ij} - a_{ij}\sum_{i=1}^{I} p_{ij} \right| \leq \varepsilon' \qquad \forall i, j, \tag{20}$$

$$\left| p_{ij} - b_{ij}\sum_{j=1}^{J} p_{ij} \right| \leq \varepsilon' \qquad \forall i, j, \tag{21}$$

$$\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1 \tag{22}$$

and

$$p_{ij} \geq 0 \ \forall i, j. \tag{23}$$

Notice that the constraints (20)–(23) are linear in the $IJ + 1$ variables p_{ij}; $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J$ and ε'. So, to determine the degree of compatibility we merely minimize the objective function

$$f(P, \varepsilon') = \varepsilon'$$

subject to (20)–(23). Any standard linear programming algorithm can be used to resolve this issue. If the minimum value of $f(P, \varepsilon')$ is zero, then A and B are compatible and the choice of P which yields the minimum value of the objective function will be the compatible joint distribution matrix. More details and alternative formulations of near compatibility may be found in Arnold, Castillo and Sarabia (1999, pages 30, 43 and 362). An alternative approach to the near compatibility problem is one in which Kulback–Leibler information distance is used to quantify the discrepancy between the conditional distributions of P and the corresponding conditional distributions given by A and B. In this case, the most nearly compatible P is found by a variation of the iterative proportional fitting algorithm much used in the study of contingency tables. For details consult Arnold and Gokhale (1994).

5. ANIL BHATTACHARYYA'S DISTRIBUTION

We return to the continuous case and focus now on probably the first conditionally specified family of densities to be studied in any detail. Bhattacharyya (1943) was actually interested in determining simple sufficient conditions to guarantee that a random vector (X, Y) would have a classical bivariate normal density. Along the way he identified a large class of joint densities with all conditional densities of the Gaussian form. He may have realized that he had in fact identified the class of *all* joint densities with normal conditionals, but he did not make that claim. His proof involved an assumption of differentiability of the densities but that assumption can easily be sidestepped. The following development is very close in spirit to that in Bhattacharyya's paper, but credit for the first carefully enunciated specification of the class of all bivariate densities with normal conditionals must be given to Castillo and Galambos (1989).

We thus are seeking all joint densities $f_{X,Y}(x, y)$ with support $\mathbb{R}^2$ such that every conditional density of X given $Y = y$ is normal with mean $\mu_1(y)$ and variance $\sigma_1^2(y)$ (which may depend on y) and every conditional density of Y given $X = x$ is normal with mean $\mu_2(x)$ and variance $\sigma_2^2(x)$ (which may depend on x). Denoting the marginal densities of X and Y by $f_X(x)$ and $f_Y(y)$, respectively, and writing the joint density of (X, Y) as a product of a marginal and a conditional density in both possible ways, yields the following functional equation:

$$\frac{f_X(x)}{\sigma_2(x)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_2(x)}{\sigma_2(x)}\right)^2\right\} = \frac{f_Y(y)}{\sigma_1(y)\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_1(y)}{\sigma_1(y)}\right)^2\right\} \tag{24}$$

We must solve (24) for $\mu_1(y), \sigma_1(y), \mu_2(x)$ and $\sigma_2(x)$ [after which $f_X(x)$ and $f_Y(y)$ will be readily obtained]. It will simplify matters if we define

$$\theta_1(y) = [\sigma_1(y)]^{-2},$$
$$\theta_2(y) = -2\mu_1(y)/\sigma_1(y),$$
$$\tau_1(x) = [\sigma_2(x)]^{-2},$$
$$\tau_2(x) = -2\mu_2(x)/\sigma_2(x).$$

Using this notation and equating (-2) times the logarithm of each side of (24) yields

$$h_1(x) + \tau_1(x)y^2 + \tau_2(x)y = h_2(y) + \theta_1(y)x^2 + \theta_2(y)x, \tag{25}$$

where $h_1(x)$ and $h_2(y)$ have been constructed by gathering all the terms involving x alone and y alone, respectively. However (25), a functional equation involving six unknown functions, is an example of a Stephanos–Levi-Civita–Suto equation. Its solution is given by the following theorem which is readily proved by differentiating. Convenient references for this and other functional equations are Aczél (1966), and Castillo and Ruiz-Cobo (1992).

THEOREM 2 [Stephanos (1904); Levi-Civita (1913); Suto (1914)]. *All solutions of the equation*

$$\sum_{i=1}^{r} f_i(x)\phi_i(y) = \sum_{j=1}^{s} g_j(y)\psi_j(x), \quad x \in S(X),\ y \in S(Y), \tag{26}$$

where $\{\phi_i\}_{i=1}^r$ *and* $\{\psi_j\}_{j=1}^s$ *are given systems of linearly independent functions, are of the form*

$$\underline{f}(x) = C\underline{\phi}(x)$$

and

$$\underline{g}(y) = D\underline{\phi}(y),$$

where $D = C'$.

Applying this result to (25) [where $\phi_1(y) = 1$, $\phi_2(y) = y$, $\phi_3(y) = y^2$, $\psi_1(x) = 1$, $\psi_2(x) = x$ and $\psi_3(x) = x^2$] we arrive at the conclusion that $h_1(x)$, $\tau_1(x)$ and $\tau_2(x)$ are quadratic functions of x and $h_2(y)$, $\theta_1(y)$ and $\theta_2(y)$ are quadratic functions of y, with interrelated coefficients (since $D = C'$ in Theorem 2).

Finally, introducing a parametrization which extends naturally to more general exponential family cases, we find that the totality of bivariate densities with normal conditionals are those of the form

$$f_{X,Y}(x, y) = \exp\left\{(1, x, x^2)\begin{pmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{pmatrix}\begin{pmatrix} 1 \\ y \\ y^2 \end{pmatrix}\right\} \tag{27}$$

subject to the constraint that the m_{ij} be chosen such that (27) is integrable.

The conditional expectations and variances are

$$E[Y|x] = -\frac{m_{01} + m_{11}x + m_{21}x^2}{2(m_{02} + m_{12}x + m_{22}x^2)}, \tag{28}$$

$$\mathrm{Var}[Y|x] = -\frac{1}{2(m_{02} + m_{12}x + m_{22}x^2)}, \tag{29}$$

$$E[X|y] = -\frac{m_{10} + m_{11}y + m_{12}y^2}{2(m_{20} + m_{21}y + m_{22}y^2)}, \tag{30}$$

$$\mathrm{Var}[X|y] = -\frac{1}{2(m_{20} + m_{21}y + m_{22}y^2)}. \tag{31}$$

We call distributions with densities of the form (27) normal conditionals distributions. Note that (27) is an eight-parameter family of densities. The coefficient m_{00} is a normalizing constant that is determined by the other m_{ij}'s and the requirement that the density integrates to 1.

Sufficient conditions for integrability of (27) are that the m_{ij}'s satisfy one of the following two sets of conditions:

$$\text{(I)}\quad m_{22}=m_{21}=m_{12}=0,\ m_{20}<0,\quad m_{02}<0 \quad\text{and}\quad m_{11}^2<4m_{02}m_{20}, \tag{32}$$

$$\text{(II)}\quad m_{22}<0, 4m_{22}m_{02}>m_{12}^2,\quad 4m_{22}m_{20}>m_{21}^2. \tag{33}$$

If (32) holds, we encounter classical bivariate normal densities. If (33) holds, we encounter non-Gaussian densities with normal conditionals.

Densities of the form (27) with m_{ij}'s satisfying (33) are markedly different from classical bivariate normal densities. They have nonnormal marginal densities. Their regression functions are either constant or nonlinear, and each regression function is bounded [see (28) and (30)]. The conditional variance functions are nonconstant [but bounded, see (29) and (31)]. The fact that the regression functions are nonlinear means that they can intersect more than once. A consequence of this phenomenon is the fact that if the m_{ij}'s satisfy (33) we may encounter bimodal (Gelman and Meng, 1991) or even trimodal (Arnold, Castillo, Sarabia and González-Vega, 2000) densities. Figures 4 and 5 show the density and contour plot of a representative bimodal density with normal conditionals. It corresponds to the parameter values

$$m_{10}=4,\quad m_{20}=-\tfrac{1}{2},\quad m_{01}=4,\quad m_{11}=0,$$
$$m_{21}=0,\quad m_{02}=-\tfrac{1}{2},\quad m_{12}=0,\quad m_{22}=-\tfrac{1}{2}.$$

[a configuration of values that satisfies (33)].

The unexpectedly bimodal marginal densities for this density are shown as projections in Figure 4. The corresponding densities of $X+Y$ and $X-Y$ are displayed in Figure 6. The figure confirms the suggestion in Section 1 that normal conditional distributions might not guarantee unimodal densities for $X+Y$ and $X-Y$.

In summary, as remarked in Section 1, the property of normal conditionals is to be found associated with some rather unusual joint densities.

REMARK 1. There is a considerable body of literature dealing with the problem of characterizing distributions by means of conditional moments rather than conditional densities. For example, we might be interested in identifying all distributions with linear regression functions and constant conditional variances (i.e., Gaussian conditional structure) (see, e.g., Arnold and Wesolowski, 1996). A related result, familiar to many in a Bayesian formulation, is that in higher dimensions linear regressions are sufficient to characterize normality (see, e.g., Rao, 1976; Goel and DeGroot, 1980). Characterizations involving one conditional density and the other regression function have also received attention (see, e.g., Wesolowski (1995); Arnold, Castillo and Sarabia, 1999, Chapter 7). Extensions to more abstract spaces are possible (see, e.g., Bischoff and Fieger, 1991).

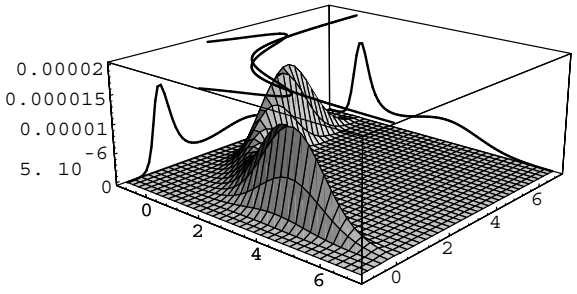

FIG. 4. *Example of a normal conditionals density with two modes showing its regression lines and its marginal densities.*

6. CONDITIONALS IN EXPONENTIAL FAMILIES

Instead of postulating that the conditional densities be normal ones, it is of interest to consider cases where quite arbitrary exponential families are playing the role of conditional densities. The main result in this section may be found in Arnold and Strauss

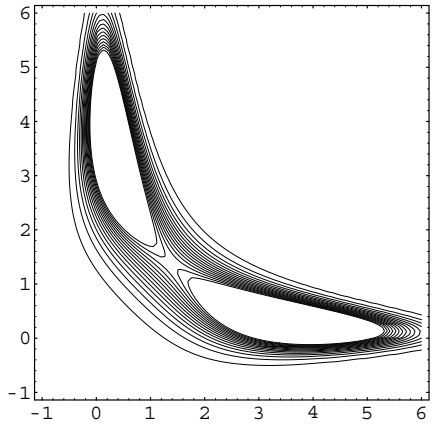

FIG. 5. *Contour plot of the normal conditionals density in Figure* 4 (*the mask of Zorro*?).

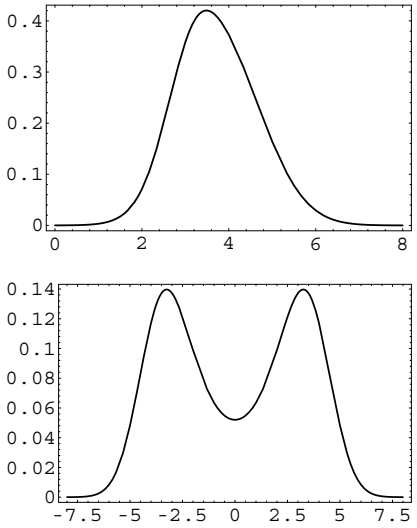

FIG. 6. *Densities of $X+Y$ (above) and $X-Y$ (below).*

(1988a,b). Analogous results in a specified stochastic process setting were discussed by Besag (1974).

DEFINITION 1 (Exponential family of densities). An l_1-parameter family of densities $\{f_1(x;\theta)\colon \theta\in\Theta\}$ with respect to μ_1 on $S(X)$ of the form

$$f_1(x;\underline{\theta}) = r_1(x)\beta_1(\underline{\theta})\exp\left\{\sum_{i=1}^{l_1}\theta_i q_{1i}(x)\right\}. \tag{34}$$

is called an exponential family of densities.

In this definition $\Theta\subset\mathbb{R}^{l_1}$ is usually the natural parameter space [all $\underline{\theta}$'s such that (34) is integrable], and the $q_{1i}(x)$'s are assumed to be linearly independent. In applications, μ_1 is often Lebesgue measure or counting measure.

In addition to the family of densities (34) we will consider a second (possibly distinct, possibly the same) l_2-parameter exponential family of densities with respect to μ_2 on $S(Y)$ given by

$$f_2(y;\underline{\tau}) = r_2(y)\beta_2(\underline{\tau})\exp\left\{\sum_{j=1}^{l_2}\tau_j q_{2j}(y)\right\}. \tag{35}$$

The class of all bivariate densities whose conditionals are in the families (34) and (35) is identified in the following theorem.

THEOREM 3 (Conditionals in exponential families). *Let $f(x,y)$ be a bivariate density whose conditional densities satisfy*

$$f(x|y) = f_1(x;\underline{\theta}(y)) \tag{36}$$

and

$$f(y|x) = f_2(y;\underline{\tau}(x)) \tag{37}$$

for some function $\underline{\theta}(y)$ and $\underline{\tau}(x)$ where f_1 and f_2 are defined in (34) *and* (35)*. It follows that $f(x,y)$ is of the form*

$$f(x,y) = r_1(x)r_2(y)\exp\{\underline{q}^{(1)}(x)M\underline{q}^{(2)}(y)'\}, \tag{38}$$

in which

$$\underline{q}^{(1)}(x) = (q_{10}(x), q_{11}(x), q_{12}(x), \ldots, q_{1l_1}(x)),$$

$$\underline{q}^{(2)}(y) = (q_{20}(y), q_{21}(y), q_{22}(y), \ldots, q_{2l_2}(y)),$$

where $q_{10}(x) = q_{20}(y) \equiv 1$ and M is a matrix of parameters of appropriate dimensions [*i.e.,* $(l_1+1)\times(l_2+1)$] *subject to the requirement that*

$$\int_{S(X)}\int_{S(Y)} f(x,y)\,d\mu_1(x)\,d\mu_2(y) = 1.$$

For convenience we can partition the matrix M as follows:

$$M = \left(\begin{array}{c|ccc} m_{00} & m_{01} & \cdots & m_{0l_2} \\ \hline m_{10} & & & \\ \cdots & & \widetilde{M} & \\ m_{l_1 0} & & & \end{array}\right). \tag{39}$$

Note that the case of independence is included; it corresponds to the choice $\widetilde{M}\equiv 0$.

PROOF. Consider a joint density with conditionals in the given exponential families. Denote the marginal densities by $g(x)$, $x\in S(X) = \{x\colon r_1(x) > 0\}$ and $h(y)$, $y\in S(Y) = \{y\colon r_2(y) > 0\}$, respectively. Write the joint density as a product of a marginal and a conditional density in two ways to obtain the relation

$$\begin{aligned} &r_1(x)r_2(y)\exp\left[\sum_{j=0}^{l_2}\tau_j(x)q_{2j}(y)\right] \\ &\quad = r_1(x)r_2(y)\exp\left[\sum_{i=0}^{l_1}\theta_i(y)q_{1i}(x)\right], \end{aligned} \tag{40}$$

where we have defined

$$\tau_0(x) = \log[g(x)\beta_2(\underline{\tau}(x))/r_1(x)],$$

$$\theta_0(y) = \log[h(y)\beta_1(\underline{\theta}(y))/r_2(y)],$$

Cancelling $r_1(x)r_2(y)$ from both sides of (40) we reduce to an equation whose solution is given directly by Theorem 2.

Of course Bhattacharyya's normal conditionals density is a fine example of a density of the form (38). The exponential conditionals density (discussed in Arnold and Strauss, 1988a) is another. For it we have

$$\begin{aligned} &l_1 = l_2 = 1,\ r_1(t) = r_2(t) = I(t>0) \\ &\quad\text{and}\quad q_{11}(t) = q_{21}(t) = -t. \end{aligned} \tag{41}$$

The resulting densities are of the form

$$f(x, y) = \exp(m_{00} - m_{01}x - m_{01}y + m_{11}xy) \times I(x > 0, y > 0).$$

For convergence we must have $m_{10} > 0$, $m_{01} > 0$ and $m_{11} \leq 0$. Densities of the form (41) always have nonpositive correlation.

An early reference which includes discussion of densities with Beta conditionals is James (1975).

As a discrete example, we might consider the class of joint densities for (X, Y) with support $\{0, 1, 2, \ldots\}^2$ such that X given $Y = y$ is a Poisson distribution for every y, and Y given $X = x$ also has a Poisson distribution for every x. This will be of the form (38) with densities with respect to counting measure on $\{0, 1, 2, \ldots\}^2$.

The specific form of such joint densities can be written in reparametrized form as

$$f_{X,Y}(x, y) = k_p(\lambda_1, \lambda_2, \lambda_3)\lambda_1^x \lambda_2^y \lambda_3^{xy}/x!y!, \quad x = 0, 1, 2, \ldots, \ y = 0, 1, 2, \ldots, \tag{42}$$

For this to be a proper joint density we must have $\lambda_1 > 0$, $\lambda_2 > 0$ and $0 < \lambda_3 \leq 1$. This Poisson conditionals distribution is also known as Obrechkoff's distribution (Obrechkoff, 1963).

Other discrete distributions with conditionals in exponential families can be defined. Arnold and Strauss (1988a), for example, describe geometric and binomial examples. Joe and Liu (1996) discuss conditionally specified logistic regression models.

A final example, of interest in the Bayesian context, to be discussed in Section 9, is the class of distributions with conditional densities of X given $Y = y$ of the normal form for every y and the conditional density of Y given $X = x$ being of the gamma form for every x. This is an example with conditionals in exponential families and so is covered by Theorem 3.

In this case we have

$$r_1(x) = 1, \qquad r_2(y) = y^{-1}I(y > 0),$$
$$q_{11}(x) = x, \qquad q_{21}(y) = -y,$$
$$q_{12}(x) = x^2, \qquad q_{22}(y) = \log y$$

and the joint density is given by

$$f(x, y) = y^{-1} \exp\left\{(1, x, x^2)M \begin{pmatrix} 1 \\ -y \\ \log y \end{pmatrix}\right\} \times I(x \in \mathbb{R}, y > 0). \tag{43}$$

There are some constraints on the m_{ij}'s in (43) needed to ensure integrability (for details on this, see Castillo and Galambos, 1989).

7. MULTIVARIATE EXTENSIONS

Although the material in Sections 2–4, and 6 was written assuming that the random variables X and Y were one-dimensional, it all remains valid if we assume that X and Y are multidimensional. Simply going back and underlining X, Y, x and y or writing them in bold face to indicate vector variables and vectors instead of real variables and real numbers is all that is required. In fact X and Y could take on values in quite abstract spaces and our results will still be valid.

EXAMPLE 5 (Logistic regression). An interesting class of $k+1$-dimensional distributions with certain conditionals in exponential families involves logistic regression models as discussed in Arnold and Press (1989). It is not unusual to assume that a binary response Y (= 0 or 1) is related to several background variables $X_1, X_2, \ldots, X_k$. We thus have

$$P(Y = 1 | X_1 = x_1, \ldots, X_k = x_k) = \left(1 + \exp - \left[\beta_0 + \sum_{j=1}^{p} \beta_j \phi_j(x_1, \ldots, x_k)\right]\right)^{-1}, \tag{44}$$

where the β_j's are unknown constants and the ϕ_j's are known "link" functions. In addition it is sometimes assumed that given $Y = 1$, (and given $Y = 0$) the vector $(X_1, X_2, \ldots, X_k)$ has some convenient joint distribution, often multivariate normal. But if we make this assumption then we are assuming that Y given $\underline{X}$ has a Bernoulli distribution and $\underline{X}$ given Y has a multivariate normal distribution. Manifestly this is an example with conditionals in exponential families and the choice of link functions in (44) will be very restricted. In fact, a family of conditional densities for Y of the form (44) will be compatible with conditional densities for $\underline{X}$ given Y only if the conditional distributions for $\underline{X}$ are in some multiparameter exponential family whose sufficient statistics are $(\phi_1(\underline{X}), \phi_2(\underline{X}), \ldots, \phi_p(\underline{X}))$. Unless the link functions (the ϕ_j's) are quite simple in their structure, the resulting distributions might be considered to be contrived and perhaps implausible.

EXAMPLE 6 (Dose response models). Let $0 < d_1 < d_2 < \cdots < d_m$ be an ordered set of dose levels. Assume that a fixed number N_i of individuals are assigned to the dose level d_i, and let X_i be the associated number of positive observations. Let $p(d; \underline{\theta})$ be the response probability to a dose $d > 0$, where $\underline{\theta}$ is a $p \times 1$ vector, where $p \leq m$, of parameters ranging over a parameter space Θ. Then, we have

$$X_i | d_i \sim \text{Binomial}(N_i, p(d_i; \underline{\theta})), \tag{45}$$

TABLE 1
Number of beetles killed after exposure to CS_2

Dose $\log_{10}[CS_2, mg/l]$	Number of insects	Killed
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

that is,

$$l(x_i;\underline{\theta}|d_i)=\binom{N_i}{x_i}p(d_i;\underline{\theta})^{x_i}(1-p(d_i;\underline{\theta}))^{N_i-x_i}, \quad x_i=0,1,\ldots,N_i. \tag{46}$$

We consider two models:

(i) Standard logistic model: we make the assumption that

$$\text{logit } p(d;\underline{\theta}) = \theta_0 + \theta_1 d, \tag{47}$$

where logit $u = \exp(u)/(1+\exp(u))$.

(ii) Quadratic logistic model: We assume that $d_i|x_i$ follows a normal distribution. Then, similarly to the case discussed in Example 5, we get

$$\text{logit } p(d;\underline{\theta}) = \theta_0 + \theta_1 d + \theta_2 d^2. \tag{48}$$

Consider the data (see Bliss, 1935) in Table 1, that gives the number of beetles deceased after exposure to different concentrations of disulphuric carbon CS_2.

The maximum likelihood parameter estimates associated with both models are shown in Table 2.

The parameter estimates are significantly different from zero. Note, however, the improvement produced by the second model that incorporates a quadratic term, not arbitrarily selected, but selcted because of conditional normality considerations.

Motivated by certain spatial models we may consider a different extension of the results of Sections 3, 5 and 6 to k dimensions. Suppose that $\underline{X}$ is a k-dimensional random vector. For each i, we define $\underline{X}_{(i)}$ to be the $(k-1)$-dimensional random vector obtained from $\underline{X}$ by deleting X_i. The same convention will be used for real vectors. Thus $\underline{x}_{(i)}$ is $\underline{x}$ with x_i deleted.

TABLE 2
Parameter estimates of the logistic standard and quadratic regression model

Parameters	Logistic model	Standard error	Quadratic model	Standard error
θ_0	−60.72	5.18	402.20	178.21
θ_1	34.27	2.91	−487.71	201.68
θ_2	—	—	147.05	57.03
$-\log l$	18.72		14.71	

Consider conditional specifications based on the collections of conditional distributions of X_i given $\underline{X}_{(i)}$, for every i. Suppose that we have k parametric families of densities on $\mathbb{R}$ given by

$$\{f_i(x;\underline{\theta}_{(i)}):\ \underline{\theta}_{(i)} \in \Theta\}, \qquad i=1,2,\ldots,k, \tag{49}$$

where $\underline{\theta}_{(i)}$ is of dimension l_i and f_i is a density with respect to μ_i, for each i. A conditionally specified model will be one in which for certain functions $\underline{\theta}_{(i)}$: $S(\underline{X}_{(i)}) \to \Theta_i$ we have

$$f_{X_i|\underline{X}_{(i)}}(x_i|\underline{x}_{(i)}) = f_i(x_i;\underline{\theta}_{(i)}(\underline{x}_{(i)})) \tag{50}$$

for every i [cf. (6) and (7)]. If (50) is to be satisfied, then an array of functional equations must hold. [cf. (8)].

In particular, consider the case in which the families of densities (49) are exponential families, that is, if

$$f_i(x;\underline{\theta}_{(i)}) = r_i(x)\exp\left\{\sum_{j=0}^{l_i}\theta_{ij}q_{ij}(x)\right\}, \quad i=1,2,\ldots,k \tag{51}$$

[where $q_{i0}(x)=1,\ \forall\ i$]. In such a setting, the functional equations can be solved (by a straightforward extension of Theorem 2) to conclude that the joint density of $\underline{X}$ must be of the form

$$f_{\underline{X}}(\underline{x}) = \left[\prod_{i=1}^{k} r_i(x_i)\right] \times \exp\left\{\sum_{i_1=0}^{l_1}\sum_{i_2=0}^{l_2}\cdots\sum_{i_k=0}^{l_k} m_{i_1,i_2,\ldots,i_k} \times\left[\prod_{j=1}^{k} q_{ii_j}(x_j)\right]\right\}. \tag{52}$$

For example, the k-dimensional analog of Bhattacharyya's distribution is of the form

$$f_{\underline{X}}(\underline{x}) = \exp\left\{\sum_{\underline{i}\in T_k} m_{\underline{i}}\left[\prod_{j=1}^{k} x_j^{i_j}\right]\right\}, \tag{53}$$

where T_k is the set of all vectors of 0's, 1's and 2's of dimension k. Densities of the form (53) have normal conditional densities for X_i given $\underline{X}_{(i)} = \underline{x}_{(i)}$ for every $\underline{x}_{(i)}$, for every i.

The classical k-variate normal density is of course a (very) special case of (53). For it, most of the $m_{\underline{i}}$'s must be zero (any $m_{\underline{i}}$ for which $\sum_{j=1}^{k} i_j > 2$).

8. ESTIMATION

Inference from conditionally specified models is somewhat complicated because of the almost ubiquitous presence of an awkward normalizing constant in the joint density. Typically, for example, the m_{00} appearing in (39) is a complicated function of the other m_{ij}'s that can only be evaluated numerically. Put another way, we usually know the shape of the likelihood but not the factor required to make it integrate to 1. In principle, maximum likelihood estimation can be implemented using a standard optimization procedure (not by solving likelihood equations). In the case of conditionals in exponential families, this approach is reasonably viable.

However, at a price of a small loss in efficiency, there are some attractive alternatives available.

One way to avoid dealing with the normalizing constant is to base the analysis on conditional distributions. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from some conditionally specified density $f(x, y; \underline{\theta})$, $\underline{\theta} \in \Theta$. We define the pseudolikelihood estimate of $\underline{\theta}$ (see Besag, 1974, 1975) to be that value of $\underline{\theta}$ which maximizes the pseudolikelihood function defined in terms of (nice) conditional densities by

$$PL(\underline{\theta}) = \prod_{i=1}^{n} f_{X|Y}(x_i|y_i;\underline{\theta}) f_{Y|X}(y_i|x_i;\underline{\theta}). \tag{54}$$

It is not difficult (see e.g., Arnold and Strauss, 1988b) to verify that such estimates are consistent and asymptotically normal. Pseudolikelihood estimates are frequently much easier to obtain than are maximum likelihood estimates. For example, if the common density of the (X_i, Y_i)'s is

$$f(x, y) \propto \exp(-x - y - \theta xy) I(x > 0,\ y > 0) \tag{55}$$

(a density with exponential conditionals), then the pseudolikelihood estimate of θ is obtained by solving

$$\sum_{i=1}^{n} \frac{X_i}{1+\theta X_i} + \sum_{i=1}^{n} \frac{Y_i}{1+\theta Y_i} = 2\sum_{i=1}^{n} X_i Y_i. \tag{56}$$

Since the left-hand side of (56) is decreasing in θ, a simple search procedure quickly yields a solution.

EXAMPLE 7 (Normal conditionals). In this example we use the Fisher data (Fisher, 1936) to estimate the parameters of the normal conditional model by maximizing the pseudolikelihood function. We have pooled together the 50 Iris-*versicolor* data points and the 50 Iris-*virginica* data points, to get a total sample size of 100.

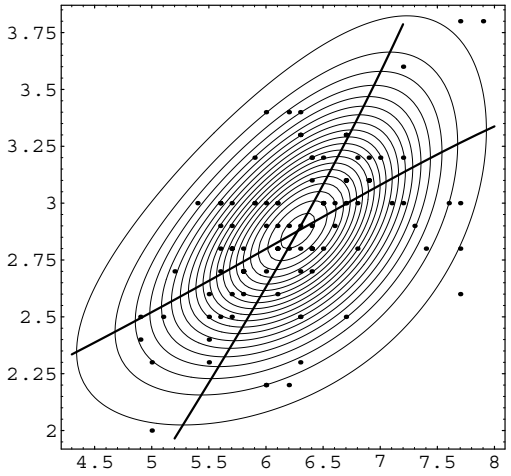

FIG. 7. *Contours and regression curves of the normal conditionals model* (27) *and data points corresponding to the Fisher Iris data example.*

The resulting model is

$$f(x, y) \propto e^{\left(\begin{smallmatrix} -1/2(-5.75\,x + 0.474\,x^2 - 100.9\,y + 0.439\,x\,y + 0.552\,x^2\,y \\ +35.82\,y^2 - 4.6\,x\,y^2 + 0.1604\,x^2\,y^2) \end{smallmatrix}\right)}.$$

See Figure 7, where the data points, the contours of the fitted probability density function and the nonlinear regression curves are shown.

EXAMPLE 8 (Skew normal conditionals). A random variable X is said to be skew-normal with parameter λ, denoted by $X \sim SN(\lambda)$, if its probability density function (pdf) is given by (see Azzalini, 1985, and Azzalini and Dalla Valle, 1996)

$$f(x;\lambda) = 2\phi(x)\Phi(\lambda x); \quad x \in \mathbb{R}, \tag{57}$$

where $\phi(x)$ and $\Phi(x)$ are the density and distribution functions of a standardized normal distribution.

Arnold, Castillo and Sarabia, (2001) study the class of bivariate distributions with skew normal conditionals, that is (X, Y) such that $X|Y = y \sim SN(\lambda_1(y))$, $\forall\, y \in \mathbb{R}$ and $Y|X = x \sim SN(\lambda_2(x))$ $\forall\, x \in \mathbb{R}$.

One of the resulting models is given by

$$f(x, y; \lambda) = 2\phi(x)\phi(y)\Phi(\lambda xy). \tag{58}$$

Such densities are not necessarily unimodal.

However, with the aim of making Model (58) more useful in practice, we introduce location and scale parameters and obtain the following more flexible family of densities:

$$f(x, y; \lambda, \underline{\mu}, \underline{\sigma}) = \frac{2}{\sigma_1\sigma_2}\phi\left(\frac{x-\mu_1}{\sigma_1}\right)\phi\left(\frac{y-\mu_2}{\sigma_2}\right) \times \Phi\left(\lambda\frac{x-\mu_1}{\sigma_1} \times \frac{y-\mu_2}{\sigma_2}\right). \tag{59}$$

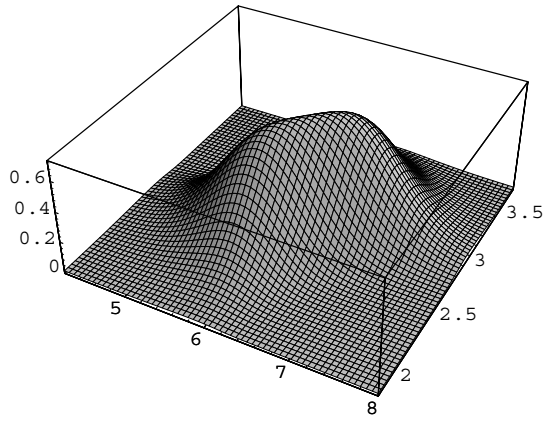

FIG. 8. *Probability density function of the skew normal conditionals model corresponding to the Fisher Iris data example.*

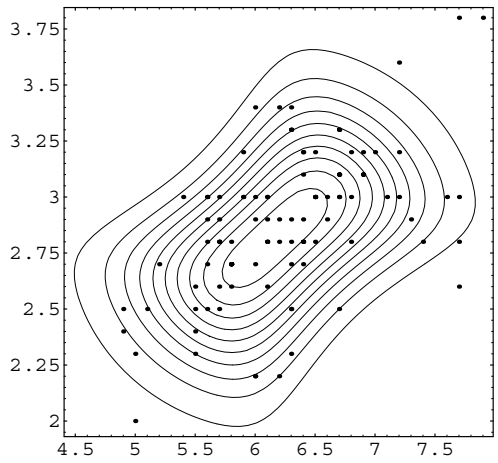

FIG. 9. *Contours of the skew normal conditionals model and data points corresponding to the Fisher Iris data example.*

The conditional densities are

$$f(x|y;\lambda,\underline{\mu},\underline{\sigma}) = \frac{2}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) \times \Phi\left(\lambda\frac{x-\mu_1}{\sigma_1}\times\frac{y-\mu_2}{\sigma_2}\right) \tag{60}$$

and

$$f(y|x;\lambda,\underline{\mu},\underline{\sigma}) = \frac{2}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right) \times \Phi\left(\lambda\frac{x-\mu_1}{\sigma_1}\times\frac{y-\mu_2}{\sigma_2}\right). \tag{61}$$

The parameters can be estimated by maximizing the pseudolikelihood function

$$L(\underline{x},\underline{y};\lambda,\underline{\mu},\underline{\sigma}) = \prod_{i=1}^{n} f(x_i|y_i;\lambda,\underline{\mu},\underline{\sigma}) f(y_i|x_i;\lambda,\underline{\mu},\underline{\sigma}). \tag{62}$$

As in the previous example, we used the Fisher data (Fisher, 1936) and pooled together the 50 Iris-*versicolor* data points and the 50 Iris-*virginica* data points, to get a total sample size of 100.

The maximum pseudolikelihood estimates are:

$$\mu_1 = 6.164; \quad \mu_2 = 2.817; \quad \sigma_1 = 0.666;$$
$$\sigma_2 = 0.335; \quad \lambda = 1.2.$$

Figures 8 and 9 show the probability density function and the contours of the fitted skew normal conditionals model, respectively.

The skew-normal conditionals example did not involve any awkward normalizing constant. For it, method of moments estimates could be readily obtained. When an awkward normalizing constant is present, a feasible approach is to use what we can call modified method of moments estimates. The standard method of moments approach to estimating a k-dimensional parameter involves judiciously selecting k functions of (X, Y), say $g_1, g_2, \ldots, g_k$ and then setting up and solving the following set of equations for $\underline{\theta}$:

$$E_{\underline{\theta}}(g_j(X,Y)) = \frac{1}{n}\sum_{i=1}^{n} g_j(X_i, Y_i), \quad j = 1, 2, \ldots, k. \tag{63}$$

In our conditionally specified settings, the expectations will typically involve the awkward normalizing constant $c(\underline{\theta})$. To avoid dealing with $c(\underline{\theta})$ we simply treat it as an extra parameter (Arnold and Strauss, 1988a) and choose an additional function g_{k+1} to allow us to augment the system (63) to now include $k+1$ equations in the, now, $k+1$ unknowns, $\theta_1, \theta_2, \ldots, \theta_k$ and c. The estimates obtained in this way for $\underline{\theta}$ are consistent asymptotically normal estimates.

For example, for a sample from the exponential conditionals density (55), denoting the normalizing constant by c we can set up the following two equations:

$$\frac{1}{n}\sum_{i=1}^{n}(X_i + Y_i) = E(X+Y) = \frac{2(c-1)}{\theta}, \tag{64}$$

$$\frac{1}{n}\sum_{i=1}^{n}(X_i Y_i) = E(XY) = \frac{(c-c^2)-\theta+2c\theta}{\theta^2}. \tag{65}$$

The solution to (64) and (65) will yield our desired estimate of θ.

A third estimation strategy is available.

We will illustrate the technique using a binomial conditionals density but the technique can be used quite generally for discrete models and, by suit-

able grouping, can be used effectively for continuous models also (see Moschopoulous and Staniswalis, 1994).

Suppose $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ are i.i.d. random variables with a common binomial-conditionals distribution. Here n_1, n_2 are fixed and known and the joint density for (X, Y) can be written in the reparametrized form

$$f_{X,Y}(x, y) = \exp[c + h(x, y) + \theta_1 x + \theta_2 y + \theta_3 xy] \times I(x \in \{0, 1, 2, \ldots, n_1\}) \times I(y \in \{0, 1, 2, \ldots, n_2\}), \tag{66}$$

where $h(x, y) = \log\binom{n_1}{x} + \log\binom{n_2}{y}$, $\theta_1 \in \mathbb{R}$, $\theta_2 \in \mathbb{R}$, $\theta_3 \in \mathbb{R}$. Of course e^c is the awkward normalizing constant. We wish to estimate $(\theta_1, \theta_2, \theta_3)$. For each possible value (i, j) of (X, Y) let N_{ij} denote the number of observations for which $X = i$ and $Y = j$. Let $\underline{N}$ denote the two-way contingency table of N_{ij}'s. The random variable $\underline{N}$ has a multinomial distribution, that is, $\underline{N} \sim$ multinomial$(n, \underline{p})$, where

$$\log p_{ij} = c + h(i, j) + \theta_1 i + \theta_2 j + \theta_3 ij. \tag{67}$$

However, the likelihood would be identical if the N_{ij}'s were independent Poisson (p_{ij}) random variables. Consequently standard Poisson regression programs can be utilized to obtain consistent asymptotically normal estimates of the θ_i's.

9. A BAYESIAN NICHE

It turns out that conditionally specified densities have potential in the role of providing convenient conjugate prior families in multiparameter cases.

Suppose that our data $\underline{X}$ has a likelihood $\{f(\underline{x}; \underline{\theta})\colon \theta \subseteq \Theta \subset \mathbb{R}^k\}$. In order to specify a suitable prior for $\underline{\theta}$, we need to describe a k-dimensional density. In some cases, this may be done conditionally as follows. Suppose that for each i, if the other parameters $\underline{\theta}_{(i)}$ were known, a convenient conjugate prior family exists for θ_i; say $\{f_i(\theta_i|\underline{\alpha}_{(i)})\colon \underline{\alpha}_{(i)} \in A_{(i)}\}$. In this notation the $\underline{\alpha}_{(i)}$'s are hyperparameters for the prior. We will then consider as a candidate family of priors for $\underline{\theta}$, those distributions such that for each i, the conditional density of θ_i given $\underline{\theta}_{(i)}$ belongs to the family f_i. It is not difficult to verify that this will yield a flexible conjugate prior family for $\underline{\theta}$ which will include as special cases the priors usually proposed for $\underline{\theta}$. Since it is a conjugate family the posterior will be in the same family and will, thus, also be conditionally specified. Simulation from the posterior will then be readily implemented using a Gibbs sampling algorithm. Note that Gibbs sampling algorithms can be implemented even when using conditional densities that are incompatible or only compatible with an improper joint density (see Hobert and Casella 1996), for some discussion of these issues). Rather than go though a general discussion of conditionally conjugate priors, we will illustrate them by considering a classical data configuration involving normal data.

TABLE 3
Adjustments in the parameters in the prior family (69)*, combined with likelihood* (68)

Parameter	Prior value	Posterior value
m_{10}	m_{10}^*	m_{10}^*
m_{20}	m_{20}^*	m_{20}^*
m_{01}	m_{01}^*	$m_{01}^* - \frac{1}{2}\sum_{i=1}^n x_i^2$
m_{02}	m_{02}^*	$m_{02}^* + n/2$
m_{11}	m_{11}^*	$m_{11}^* + \sum_{i=1}^n x_i$
m_{12}	m_{12}^*	m_{12}^*
m_{21}	m_{21}^*	$m_{21}^* - n/2$
m_{22}	m_{22}^*	m_{22}^*

Suppose that our available data consist of n i.i.d. normal random variables with mean μ and precision (=reciprocal of the variance) τ.

The likelihood is of the form

$$f_{\underline{X}}(\underline{x}; \mu, \tau) = \frac{\tau^{n/2}}{(2\pi)^{n/2}} \exp\left[-\frac{\tau}{2}\sum_{i=1}^n (x_i - \mu)^2\right]. \tag{68}$$

If τ were known, a natural conjugate prior family for μ would be the normal family. If μ were known, a natural conjugate prior family for τ would be the gamma family. This suggests that an appropriate conjugate prior family for (μ, τ) (assuming both are unknown) would be one in which μ given τ is normally distributed for each τ, and τ given μ has a gamma distribution for each μ.

Such normal-gamma conditionals distributions were discussed in Section 6. They form an eight-parameter exponential family of distributions with densities of the form

$$f(\mu, \tau) \propto \exp[m_{10}\mu + m_{20}\mu^2 + m_{12}\mu \log\tau + m_{22}\mu^2 \log\tau] \times \exp[m_{01}\tau + m_{02}\log\tau + m_{11}\mu\tau + m_{21}\mu^2\tau]. \tag{69}$$

In Table 3 we summarize the relationships between prior and posterior values of the (hyper) parameters appearing in (69), when combined with the likelihood (68).

It is evident that 4 of the m_{ij}'s [those appearing in the first factor in (69)] are not affected by the data. The other four are changed by the data. The

family (69) includes both of the usually suggested joint priors for (μ, τ). The classical prior for (μ, τ) (see, e.g., DeGroot, 1970) corresponds to the choice

$$m_{10} = m_{20} = m_{12} = m_{22} = 0. \tag{70}$$

Condition (70) enforces a peculiar dependence structure on the joint prior for (μ, τ). It is not self-evident that this will always adapt well to prior beliefs.

A second approach, advocated by those who view marginal assessment of prior beliefs to be the most viable (see, e.g., Press, 1982), assumes independent gamma and normal marginals in (70). This corresponds to initially setting

$$m_{11} = m_{12} = m_{21} = m_{22} = 0. \tag{71}$$

Here too, we might not always find the implied dependence structure (in this case independence) adapting well to our expert's prior beliefs. The more flexible full family (69) will retain the dual advantages of conjugacy and ease of simulation.

The large number of hyperparameters appearing in conditionally specified priors [such as (69)] might be cause for concern if assessment of appropriate prior values were viewed as a difficult problem. Of course many or all could be set to zero to attain varying degrees of diffuseness in the prior. It is however possible to develop quite simple assessment algorithms involving prior specification of an array of conditional moments and/or percentiles (Arnold, Castillo and Sarabia, 1999).

The conditional specification approach extends quite readily to more interesting cases, such as Behrens–Fisher problems, variance components, 2×2 contingency tables, etc. For details and references see Arnold, Castillo and Sarabia (1999).

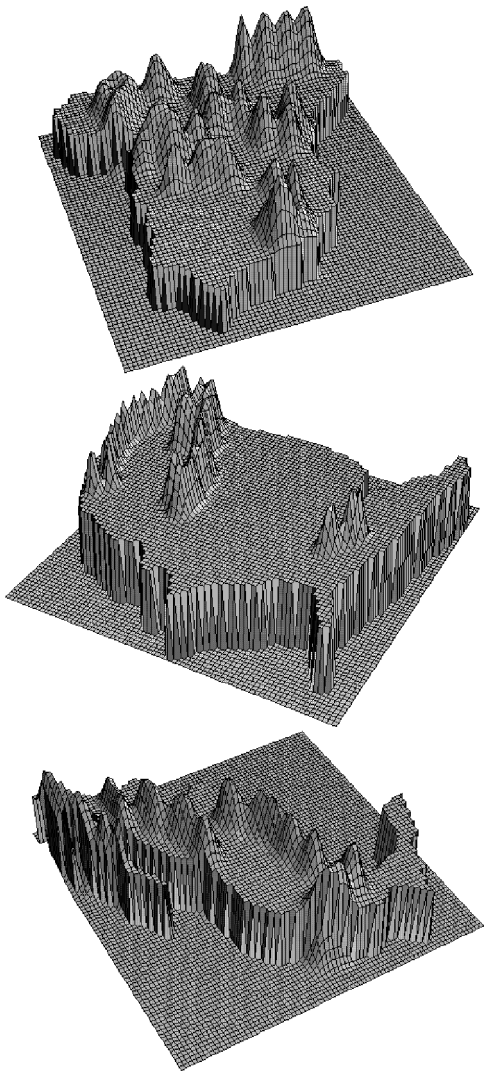

FIG. 10. *Joint densities (i.e., normalized topographical maps) of Spain, the United States and Mexico.*

10. THE PROBLEM OF MARGINAL AND CONDITIONAL SPECIFICATION

When one is dealing with the distribution of a k-dimensional random variable, there are a large number of possible ways in which it can be specified in terms of marginal and conditional densities. Gelman and Speed (1993) provide a careful discussion of what combinations of marginal and conditional densities will suffice to determine either a unique joint density or a class of compatible joint densities. For modeling purposes we are especially interested in identifying which combinations of marginals and conditionals will uniquely determine the joint density of $(X_1, \ldots, X_k)$. For example, if we are given the density of X_i given $\underline{X}_{(i)}$ for each i, those, if consistent, will determine the joint density and, under a positivity condition (or indecomposibility of a related Markov chain), they will uniquely determine the joint density of $\underline{X}$. In contrast, if we are given the conditional density of X_i given X_j for every i and j, this will fail to determine the joint density of $\underline{X}$ (provided the dimension exceeds 2). This is true because the given information can only provide information about two-dimensional marginals and not the full joint density. There are some curious configurations that sometimes work. It is obvious in two dimensions that the marginal density of X_1 and the conditional density of X_2 given X_1 will always uniquely determine the joint distribution of (X_1, X_2). It is, however, sometimes possible (see Seshadri and Patil, 1964) that the marginal distribution of X_1 and the conditional density of X_1 given X_2 (the "wrong" conditional, if you wish) will uniquely determine the joint distribution of (X_1, X_2). But it doesn't always work.

Gelman and Speed noted that knowledge of any conditional density of the form

$$\{X_i: i \in A\} \text{ given } \{X_j: j \in B\},$$

where B could be empty (corresponding to a marginal rather than a conditional density), is equivalent to knowledge of the conditional density of individual X_i's given other X_j's. Thus, in transparent notation: knowledge of $X_1, X_3|X_2, X_5$ is the same as knowing $X_1|X_2, X_5$ and $X_3|X_1, X_2, X_5$

To be always certain that given marginal and conditional information will, if consistent, determine the joint distribution, it must include the distribution of X_i given $\underline{X}_{(i)}$ for some i and, sufficient additional marginal and conditional information to determine the distribution of $\underline{X}_{(i)}$.

Careful delineation of these claims may be found in Arnold, Castillo and Sarabia (1999, Chapter 10), building on the material in Gelman and Speed (1993). Note that the necessary part of the main theorem as stated in Gelman and Speed (1993) is incorrect. A counterexample is documented in Gelman and Speed (1999). Suitable modification of the statement of this theorem is discussed in Arnold, Castillo and Sarabia (1999).

11. ENVOI

We close by reiterating our claim that conditional specification is quite natural. Cross-sectional descriptions are quite basic. Conditional specification provides a much needed augmentation in the flexibility of parametric multivariate models. As in many branches of statistics, life is smoothest when dealing with exponential families. The concept of conditional specification does provide us with some new (or old as in the case of the Bhattacharyya density!) alternatives to the customary tool box of models available in classical and Bayesian statistical scenarios. It doesn't come with a money back guarantee, but we recommend you give it a try.

ACKNOWLEDGMENT

We are grateful for constructive editorial suggestions which have improved and clarified the material in this paper.

REFERENCES

ACZÉL, J. (1966). *Lectures on Functional Equations and their Applications*. Academic Press, New York.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1999). *Conditional Specification of Statistical Models*. Springer, New York.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001). Conditionally specified multivariate skewed distributions. *Sankhyā*. To appear.

ARNOLD, B. C., CASTILLO, E., SARABIA, J. M. and GONZÁLEZ-VEGA, L. (2000). Multiple modes in densities with normal conditionals. *Statist. Probab. Lett.* **49** 355–363.

ARNOLD, B. C. and GOKHALE, D. V. (1994). On uniform marginal representations of contingency tables. *Statist. Probab. Lett.* **21** 311–316.

ARNOLD, B. C. and PRESS, S. J. (1989). Compatible conditional distributions. *J. Amer. Statist. Assoc.* **84** 152–156.

ARNOLD, B. C. and STRAUSS, D. (1988a). Bivariate distributions with exponential conditionals. *J. Amer. Statist. Assoc.* **83** 522–527.

ARNOLD, B. C. and STRAUSS, D. (1988b). Pseudolikelihood estimation. *Sankhyā, Ser. B* **53** 233–243.

ARNOLD, B. C. and WESOLOWSKI, J. (1996). Multivariate distributions with Gaussian conditional structure. *Stochastic Processes and Functional Analysis. Lecture Notes in Pure and Applied Mathematics* **186** 45–59. Dekker, New York.

AZZALINI, A. (1985). A class of distributions which includes the normal ones. *Scand. J. Statist.* **12** 171–178.

AZZALINI, A. and DALLA VALLE, A. (1996). The multivariate skew-normal distribution. *Biometrika* **83** 715–726.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. E. (1975). Statistical analysis of nonlattice data. *Statistician* **24** 179–196.

BHATTACHARYYA, A. (1943). On some sets of sufficient conditions leading to the normal bivariate distribution. *Sankhyā* **6** 399–406.

BISCHOFF, W. and FIEGER, W. (1991). Characterization of the multivariate normal distribution by conditional normal distributions. *Metrika* **38** 239–248.

BLISS, C. I. (1935). The calculation of the dosage-mortality. *Ann. Appl. Biol.* **22** 134–167.

CASTILLO, E., COBO, A. JUBETE, F. and PRUNEDA, R. E. (1999). *Orthogonal Sets and Polar Methods in Linear Algebra: Applications to Matrix Calculations, Systems of Equations and Inequalities, and Linear Programming*. Wiley, New York.

CASTILLO, E. and GALAMBOS, J. (1987). Lifetime regression models based on a functional equation of physical nature. *J. Appl. Probab.* **24** 160–169.

CASTILLO, E. and GALAMBOS, J. (1989). Conditional distributions and the bivariate normal distribution. *Metrika* **36** 209–214.

CASTILLO, E. and RUIZ-COBO, R. (1992). *Functional Equations in Science and Engineering*. Dekker, New York.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7** 179–188.

GELMAN, A. and MENG, X. L. (1991). A note on bivariate distributions that are conditionally normal. *Amer. Statist.* **45** 125–126.

GELMAN, A. and SPEED, T. P. (1993). Characterizing a joint probability distribution by conditionals. *J. Roy Statist. Soc. Ser. B* **55** 185–188.

GELMAN, A. and SPEED, T. P. (1999). Corrigendum: Characterizing a joint probability distribution by conditionals. *J. Roy. Statist. Soc. Ser. B* **61** 483.

GOEL, P. K. and DEGROOT, M. H. (1980). Only normal distributions have linear posterior expectations in linear regression *J. Amer. Statist. Assoc.* **75** 895–900.

GOURIEROUX, C. and MONTFORT, A. (1979). On the characterization of a joint probability distribution by conditional distributions. *J. Econometrics* **10** 115–118.

HOBERT, J. P. and CASELLA, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *J. Amer. Statist. Assoc.* **91** 1461–1473.

JAMES, I. R. (1975). Multivariate distributions which have beta conditional distributions. *J. Amer. Statist. Assoc.* **70** 681–684.

JOE, H. and LIU, Y. (1996). A model for a multivariate binary response with covariates based on compatible conditionally specifed logistic regressions. *Statist. Probab. Lett.* **31** 113–120.

KOTZ, S., BALAKRISHNAN, N. and JOHNSON, N. L. (2000). *Continuous Multivariate Distributions 1: Models and Applications*, 2nd ed. Wiley, New York.

LEVI-CIVITA, T. (1913). Sulle funzioni che ammettono una formula d'addizione del tipo $f(x + y) = \sum_{i=1}^{n} X_i(x)Y_i(y)$. *Atti della Accademia Nazionale dei Lincei, Rendiconti* **5** 181–183.

MOSCHOPOULOS, P. and STANISWALLIS, J. G. (1994). Estimation given conditionals from an exponential family. *Amer. Statist.* **48** 271–275.

OBRECHKOFF, N. (1963). *Theory of Probability*. Nauka i Izkustvo, Sofia.

PRESS, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Krieger, Melbourne, FL.

RAO, C. R. (1976). Characterization of prior distributions and solution to a compound decision problem. *Ann. Statist.* **4** 823–835.

SESHADRI, V. and PATIL, G. P. (1964). A characterization of a bivariate distribution by the marginal and the conditional distributions of the same component. *Ann. Inst. Statist. Mathe.* **15** 215–221.

STEPHANOS, C. (1904). Sur une categorie d'equations fonctionnelles. *Rend. Circ. Mat. Palermo* **18** 360–362.

SUTO, O. (1914). Studies on some functional equations. *Tohuku Math. J.* **6** 1–15.

WESOLOWSKI, J. (1995), Bivariate discrete measures via a power series conditional distribution and a regression function. *J. Multivariate Analy.* **55** 219–229.

Comment

Julian Besag

I am grateful for the opportunity to comment on this paper and on the role of conditionally specified distributions. My contribution will focus on the multivariate rather than the bivariate problem. I think my 1974 paper, mentioned in passing by the present authors, was the first to exploit the general idea of constructing a joint density $f(x)$ for a random vector $X = (X_1, \ldots, X_n)$ from its local characteristics or, as they are now known, full conditionals $f_i(x_i|x_{-i})$, where x_{-i} denotes the values taken by all the X_j's other than X_i. This paper owes a great deal to Maurice Bartlett, with whom I was privileged to work from 1968–1969 (see also Bartlett, 1967), to Hammersley and Clifford (1971; see also Clifford 1990) and, for stationary Gaussian lattice systems, to Lévy (1948).

In spatial statistics, the conditional probability approach is motivated by applications of the following type. Consider an agricultural experiment in which different varieties of a crop occur over a rectangular array of plots according to a particular design, with rather limited replication. The plots are harvested, their yields are measured and these observations are used to make comparisons between the effects of the different varieties. It is easy to accommodate standard fixed effects due to management practice but usually the yields are also heavily influenced by variation in fertility over the experimental region. Direct measurements of soil properties are not made in general and this makes estimation difficult. The Fisherian solution uses randomization, within the block constraints of the design, to induce a corresponding reference distribution. Alternatively, one may represent the true fertility x_i in each plot i by an unknown X_i and then assign an appropriately flexible spatial distribution to the X_i's, whether from a frequentist or a Bayesian viewpoint. The intention is that accuracy and precision of estimates will improve, without the need for very intricate designs that plant breeders are often unwilling to implement. Note that, because the main focus is the comparison of variety effects and not the estimation of the x_i's and because of the replication, the conclusions from the analysis should be relatively insensitive to a quite wide range of spatial distributions. Spatial formulations were first proposed in the 1920's by the distinguished Greek agronomist, J. S. Papadakis, but received scant attention from statisticians until the 1980s, apart from Bartlett (1938, 1978). However, the past twenty years has seen a flurry of activity and now overtly spatial analyses are used quite widely; for example, in some 5000 field experiments annually in Australia.

To see the relevance of conditional specifications to field experiments, suppose that $f(x)$ is a postulated joint density for the plot fertilities X_i. Then

Julian Besag is Professor, Department of Statistics, Box 354322, University of Washington, Seattle, Washington 98195 (*e-mail: julian@stat.washington.edu*).

the full conditional $f_i(x_i|x_{-i}) \propto f(x)$ is a rather natural object to examine because of the spatial context. This contrasts with time series analysis, where the obvious conditioning set for any particular random variable consists of its predecessors, a meaningless concept in purely spatial settings. It is now a short step to suggest that *formulation* of $f(x)$ should be in made in terms of its full conditionals. That is, for each plot i, one imagines knowing the fertilities x_{-i} in all other plots and attaching a corresponding conditional distribution to X_i. Note that this involves a form of interpolation rather than extrapolation. A starting point is to assume that each $f_i(x_i|x_{-i})$ is Gaussian with a mean that is a linear combination of the fertilities in plots that are closest to i. If outliers or jumps in fertility need to be accommodated, the Gaussian form can be replaced by one based on spatial medians or by a hierarchical–t. The interested reader may consult Besag and Higdon (1999) for further details and the analysis of several awkward datasets from a Bayesian perspective. Of course, it must be ensured that the formulation knits together to produce a genuine $f(x)$. Note that it is also possible to mix and match conditional and joint assumptions, as in Besag and Kooperberg (1995), and this is often my preferred approach.

The above topic, in an important area of applied science, I hope provides some motivation for considering conditional specifications. I could instead have chosen examples from geographical epidemiology, statistical ecology, texture analysis, semi conductor manufacturing, computerized tomography, synthetic magnetic resonance imaging or social networks, to all of which conditional specifications have been applied, with varying degrees of success. I could even have discussed the analysis of higher dimensional contingency tables, where conditional probability formulations provide an alternative route to hierarchical models; or the example, pointed out by Stephen Stigler, in which Francis Galton adopted the approach as an aid to the statistical analysis of fingerprint data in the 1890s! My point is that there is no need to resort to contrived examples of the type described by the present authors in their Section 1. Indeed, their reduction of an inherently spatial problem to a bivariate distribution seems strange from any viewpoint.

I would like now to return to my 1974 paper. Although this is set in the context of spatial statistics, where conditionally specified distributions are known as Markov (random) fields, the mathematics is rather general and draws on the Hammersley–Clifford theorem. This establishes a correspondence between such fields and Gibbs distributions in statistical physics but is not tied to lattice systems or even to Euclidean space. Indeed, it is embarrassing that my own attempts to solve the problem had been far too specific and merely led to an extremely limited result, obtained by some turgid mathematics! My 1974 paper is more enlightened and, in particular, Section 4 examines the consequences of assuming that only pairwise interactions exist between the n random variables and also that the full conditionals belong to an exponential family. The first of these restrictions is of course vacuous for a bivariate distribution and, as regards the second, the development in Besag (1974) extends to the more general definition of an exponential family adopted in Section 6 of the present paper. It is rather misleading of the authors to suggest that the 1974 results arise "in a specified stochastic process setting," because not only is the mathematics quite separate from any spatial context but also I emphasize the distinction between a space–time formulation (or "process") and a purely spatial one (or "scheme"). Although the latter can be interpreted as the instantaneous cross-section of a space–time process, assuming temporal stationarity (e.g., Besag, 1977), this is generally unconvincing and is of course untestable from purely static data. I contend that conditionally specified distributions usually arise from thought processes rather than physical ones, and this may fit more comfortably into a Bayesian rather than a frequentist paradigm.

Besag (1974, Section 4) also discusses special cases of distributions with pairwise interactions and exponential family full conditionals, including so-called auto-logistic, auto-binomial, auto-Poisson, auto-exponential, auto-gamma and auto-Normal schemes, with the obvious connotation that an auto-binomial distribution has binomial/full conditionals and so on. The auto-exponential and auto-Poisson are the generalizations of (41) and (42) to n variables. Also, it is remarked in the 1974 paper that the Poisson, exponential and gamma versions are probably of little practical interest because of the parameter restrictions (though this has not prevented their occasional abuse). In brief, the references to specific distributions in Section 6 of the present paper seem very carefully selected!

The auto-logistic scheme can also be thought of as a truncated Bahadur expansion for the distribution of multivariate binary variables, without recourse to conditional probability. It has been rediscovered as the Boltzmann distribution in computer science and as the quadratic exponential in statistics. It is used quite widely in statistical ecology and in the study of familial diseases, where it often takes the form of a so-called auto-logistic regression model,

sometimes claimed to be a generalization of the autologistic but in fact a special case. Another special case is the celebrated Ising model in statistical physics. Although this is derived classically from thermodynamic principles, it is remarkable that the Ising model follows necessarily as the very simplest non-trivial binary Markov random field on any finite d-dimensional cubic lattice. The Ising model has been used with some success as a prior distribution for object against background in low-level Bayesian image reconstruction but is not suitable for more complicated tasks, which require cumbersome higher-order Markov random fields; see Tjelmeland and Besag (1998). A superior approach is based on Ulf Grenander's deformable templates.

In Section 8 of their paper, the authors resurrect maximum pseudolikelihood estimation. Originally, this was devised for Markov random fields, for which the normalizing constant in the likelihood function is usually horrendous. Quite generally, the pseudolikelihood is defined as the product of the full conditionals for the observed data (Besag, 1975) and, in particular, for a random sample from a bivariate distribution, this indeed leads to equation (54). Pseudolikelihood also extends to Markov point processes (Besag, 1978), where it is defined in terms of the corresponding Papangelou conditional intensities. For recent work, including central limit theorems, see, for example, Comets and Janzura (1998), Baddeley and Turner (2000) and Baddeley (2001). Note that asymptotics for spatial systems are much more difficult than for the repeated bivariate or multivariate samples of Section 8. Also Baddeley (2000) gives an interesting discussion of the relationship between pseudolikelihood and other methods of estimation in a general setting. My own view is that the technique is really a creature of the 1970's and 1980s and I am surprised to see it recommended in the computer age, all the more so in the rather undemanding context of bivariate distributions.

Despite its consistency, maximum pseudolikelihood performs poorly in most spatial systems with strong interactions and it is easy to find examples in the literature where the estimates are of questionable value. This is also true in the analysis of social networks, formulated conditionally in terms of a Markov assumption due to Frank and Strauss (1986) that has some intuitive appeal. It is not always the case that the price paid for the simplicity of pseudolikelihood is merely "a small loss in efficiency," even for bivariate distributions. Perhaps the main role for the method these days is to provide initial approximations for iterative maximum likelihood algorithms, whether deterministic or Monte Carlo. An extreme example in which pseudolikelihood is not even consistent has been pointed out by Gareth Roberts. Consider a random walk $X_1, \ldots, X_n$, seeded by $x_0 = 0$ and with independent $N(\mu, 1)$ increments. Then the likelihood and pseudolikelihood estimators of μ are X_n/n and $X_n - X_{n-1}$, respectively!

Finally, almost all Markov chain Monte Carlo algorithms, whether Gibbs sampler or otherwise, are driven by univariate full conditionals, so that conditional specifications arise automatically. Not surprisingly, the original uses of the methods in statistics were for conditionally specified spatial distributions. As the authors point out in their Section 4, one can also imagine a Gibbs sampler running with *incompatible* full conditionals, a procedure that Heckerman, Chickering, Meek, Rounthwaite and Kadie (2000) refer to as a pseudo-Gibbs sampler. The analogous idea appears in Section 6 of Besag (1986) but for a greedy deterministic version of Gibbs known as iterated conditional modes. One possible application occurs in model criticism. Suppose that, in a basic formulation, the full conditionals are sufficiently interpretable that some invite possible modification without particular regard to others. For example, this happens in the basic Gaussian formulations for plot fertilities, mentioned previously. Then, in general, the new full conditionals will not be strictly compatible. Nevertheless, one can run a pseudo-Gibbs algorithm and obtain samples from the limiting distribution, presuming this exists, which should be roughly consistent with the amended full conditionals. One might use a random scan rather than a fixed one to produce a unique end result. In the research at Microsoft, Heckerman et al. (2000) are interested in synthesizing joint distributions from conditionals, derived empirically and without regard to compatibility. The theoretical properties of pseudo-Gibbs samplers are largely unknown and no doubt considerable caution must be exercised. It would be of interest to extend the methods in Section 4 beyond the bivariate case.

ACKNOWLEDGMENT

I am grateful for the support of the National Research Center for Statistics and the Environment at the University of Washington.

Comment

Andrew Gelman and T. E. Raghunathan

1. INTRODUCTION

The authors discuss conditionally specified models in probability theory and for modeling joint distributions in various applications. This theoretical structure is useful, considering that conditional models are becoming standard in many spatial applications, following Besag (1974). (Rather than attempting an exhaustive or even representative list, we shall just refer to Besag and Higdon, 1999, as a recent example with discussion.) In addition, there has been occasional discussion in the literature as to the relative merits of conditionally or jointly specified models (for example, Besag, 1974; Haslett, 1985; Ripley, 1988).

Here, however, we would like to address a different topic: the use of conditional distributions, not to model an underlying joint distribution, but for the purpose of imputing missing data. At first this might seem like an unimportant distinction—after all, imputation requires modeling (if only implicitly). However, when the fraction of missing data is not large, imputations can be reasonable even if they are not based on the correct complete-data model (see Meng, 1994; Rubin, 1996). Thus, it makes sense to consider modeling for imputation separately from modeling of underlying phenomena.

We shall refer to the example of the New York City Social Indicators Survey (Garfinkel and Meyers, 1997), where we had to impute missing responses for family income conditional on demographics and information such as whether or not anyone in the family received government welfare benefits. Conversely, if the "welfare benefits" indicator is missing, then family income is clearly a useful predictor. The whole situation was actually more complicated because the survey asked about several different sources of income, and these questions had different patterns of nonresponse.

Andrew Gelman is Professor, Department of Statistics and Director of the Quantitative Methods in Social Science Program, Columbia University, 2990 Broadway, New York, New York, 10027 (e-mail: gelman@stat.columbia.edu). T. E. Raghunathan is Professor, Department of Biostatistics and Institute for Social Research, University of Michigan, Ann Arbor, Michigan.

2. INCONSISTENT CONDITIONAL DISTRIBUTIONS

As discussed by the authors, a multivariate normal distribution has conditionals that are normal linear regressions, in which case the conditional distributions are automatically compatible. However, when any of these conditions is relaxed—that is, if data are bounded or discrete (and thus cannot be modeled as normal), or regression relationships are nonlinear or have interactions—then, in general, reasonable-seeming conditional models will not be compatible with any single joint distribution.

Nonetheless, imputations can be performed using conditional models; that is, one can start with guesses of all the missing data, then impute $x_1|x_2, x_3, \ldots, x_k$, impute $x_2|x_1, x_3, \ldots, x_k$, and so forth, looping indefinitely through all the variables. If the imputations are stochastic, this is just the notorious "inconsistent Gibbs" algorithm, for which the simulation draws never converge to a single joint distribution; rather, the distribution depends upon the order of the updating and on when the updating is stopped.

With the inconsistent Gibbs sampler, one is always afraid of reasonable-seeming conditional distributions that produce a diverging random walk; for example, if $x_1|x_2 \sim N(x_2, 1)$, and $x_2|x_1 \sim N(x_1, 1)$, then the distribution of the simulations simply diffuses out to infinity. However, in practice, with the distributions estimated from data (and using constraints or proper prior distributions when dimensions are high and data sparse), this should not happen.

A big advantage of conditional (rather than joint) modeling is that it splits a k-dimensional problem into k one-dimensional problems, each of which can be attacked flexibly. Thus, conditional imputation using k separate regression models is a popular approach, and it has recently been formalized by Raghunathan, Lepkowski, Solenberger and Van Hoewyk (2001) and implemented in SAS-compatible software (Raghunathan, Solenberger and Van Hoewyk, 1997). This particular program allows continuous variables to be modeled using normal distributions, binary variables with logistic regression, with other options for ordered and unordered discrete variables and for continuous variables with constraints. The corresponding joint posterior distribution may not exist, of course, which means that the Bayesian inference used to get uncertainties for the imputations is only uncertain. [It could, however, possibly be formalized as a Bayesian counterpart to the pseudolikelihood (Besag, 1975), in which the likelihood function is replaced by the product of conditional densities.]

Performing imputation is awkward without a joint model, and it also results in difficulties in inference for the imputation model itself (for example, how do you correctly adjust for truncation in a bounded-variable model when there is no joint distribution over which to integrate). However, the separate regressions often make more sense than joint models which either assume normality and hope for the best (Gelman, King and Liu, 1998) or mix normality with completely unstructured discrete distributions (Schafer, 1997) or mix normality (with random effects) and log-linear structures for discrete distributions (Raghunathan and Grizzle, 1995) or generalize with the t distribution (Liu, 1995). From a practical perspective, all these approaches provide useful tools, and some of the time it will make sense to go with the inconsistent, but flexible, conditional models such as described by Raghunathan et al. (2001).

One may argue that having a joint distribution in the imputation is less important than incorporating information from other variables and unique features of the data set (such as zero–nonzero features in income components, bounds, skip patterns, nonlinearity, interactions and so forth). Conditional modeling allows enormous flexibility in dealing with practical problems. We have never been able to apply the joint models to a real data set without making drastic simplifications.

However, if one is modeling some aspect of the nature, then the joint distribution has to be the end point. Specifying just the conditionals without a coherent joint distribution will not be acceptable. However, many of our applied collaborators are just as happy with conditionals such as p(Hypertension | Body Mass Index) or p(Body Mass Index | Hypertension, Socioeconomic Status), rather than p(Hypertension, Body Mass Index, Socioeconomic Status).

3. CHOICES IN SETTING UP THE IMPUTATION MODELS

We conclude with a discussion of an awkward (or perhaps promising) issue: structural features of the conditional models can affect the distributions of the imputations in ways that are not always obvious. To return to the example introduced at the end of Section 1 of this discussion, suppose we are imputing a continuous income variable y_1, and a binary indicator y_2 for welfare benefits, conditional on a set X of fully observed covariates.

We can consider two natural approaches. Perhaps simplest is a direct model where, for example, $p(y_1|y_2, X)$ is a normal distribution (perhaps a regression model on y_2, X, and the interactions of y_2 and X) and $p(y_2|y_1, X)$ is a logistic regression on y_1, X, and the interactions of y_1 and X. (For simplicity, we ignore the issues of nonnegativity and possible zero values of y_1.)

A more elaborate, and perhaps more appealing model uses hidden variables: let z_2 be a latent continuous variable, defined so that

$$y_2 = \begin{cases} 1, & \text{if } z_2 \geq 0, \\ 0, & \text{if } z_2 < 0. \end{cases} \tag{1}$$

We can then model $p(y_1, z_2|X)$ as a joint normal distribution (i.e., a multivariate regression). Compared to the direct model, this latent-variable approach has the advantage of a consistent joint distribution. And, once inference for (y_1, z_2) has been obtained, we can directly infer about y_2 using (1). In addition, this model has the conceptual appeal that z_2 can be interpreted as some sort of continuous "proclivity" for welfare, that is, only activated if it exceeds a certain threshold. In fact, the relation between z_2 and y_2 can be made stochastic if such a model would appear more realistic.

So the latent-variable model is better (except for possible computational difficulties), right? Not necessarily. A perhaps disagreeable byproduct of the latent model is that, because of the joint normality, the distributions of income among the welfare and non welfare groups—that is, the distributions $p(y_1|y_2 = 1, X)$ and $p(y_1|y_2 = 0, X)$—must necessarily overlap. In contrast, the direct model allows there to be overlap or nonoverlap, depending on the data. Thus, although the latent-variable model seems to be a generalization, it is not.

4. CONCLUSIONS

Where does this leave us in practice? Must we just choose a model and hope for the best? Fortunately, we are not completely without tools: in particular, we can use a procedure to impute missing data and then check the fit of the model to the completed dataset (Gelman, King and Liu 1998; Gelman et al., 2001). Serious problems (such as overlapping distributions for imputed data amidst nonoverlapping distributions of observed data) should show up. With checking, we should be able to notice major flaws in an imputation model. But we do not have a good sense of how general the models have to be in order to work well, and it is not clear when incompatibility of conditional distributions presents a practical problem.

As with so much of statistics, the study of conditional distributions is an area where theory has not caught up with practice.

5. ACKNOWLEDGMENT

We thank the NSF for support through Grants SES-9987748 and SES-0084368.

Comment

Harry Joe

The authors have provided a very nice introduction to topic of conditionally specified models. There is nothing in the paper that I disagree with, so this discussion is a supplement to the paper that aims to provide intuitive understanding of some of the results and some guide in choices of conditional distributions. The main topic is relating properties of conditional distributions to strength of dependence.

Recently several books (Joe, 1997; Arnold, Castillo and Sarabia, 1999; Nelsen, 1999; Kotz, Balakrishnan and Johnson, 2000) with theory for multivariate nonnormal families have been published. In the framework of given univariate margins rather than given conditional distributions, a multivariate distribution is said to be of type x if all of its univariate margins are of type x. The most commonly used multivariate distribution is multivariate normal. Other than $x =$ normal and $x =$ Poisson, there is no "natural" multivariate family with a given parametric family for the univariate margins, and a common approach has been through copulas. Because of some difficulty in the construction of multivariate copulas (in dimensions greater than or equal to 3) with nice properties, the method of conditional specified distributions is a good approach to consider. However, in this case one should pay attention to the type and range of dependence in the resulting multivariate family.

There are some unusual or surprising results for conditional distributions in exponential families. For example, only negative dependence is possible if conditional distributions are all exponential or Poisson. I will demonstrate that these unusual results occur mainly in one-parameter families. The explanation can be seen from the range of possible dispersions in the conditional distributions.

Properties to keep in mind when specifying conditional distributions are the following:

1. Positive dependence is obtained if the conditional distributions are stochastically increasing in the variables being conditioned on.
2. Negative dependence is obtained with the stochastic decreasing behavior.
3. With positive dependence, conditional distributions are generally less dispersed (say, as measured through conditional variance or conditional coefficient of variation) than the univariate margins, and conditional dispersion decreases as the amount of positive dependence increases.

These properties are illustrated with a few examples of (1), known bivariate distributions with given margins, and (2), bivariate distributions with conditional distributions in exponential families.

EXAMPLE 9. (X, Y) has the bivariate Poisson distribution if the stochastic representation

$$(X, Y) \stackrel{d}{=} (Z_{12} + Z_1, Z_{12} + Z_2),$$

where Z_1, Z_2, Z_{12} are independent Poisson random variables with means $\lambda_1, \lambda_2, \lambda_{12}$, respectively (a Poisson random variable with mean zero is equivalent to the degenerate random variable at zero). The conditional distribution of Y given $X = x$ corresponds to a convolution of a Binomial$(x, \lambda_{12}/[\lambda_1 + \lambda_{12}])$ random variable and a Poisson(λ_2) random variable. Hence

$$\begin{aligned} \mathrm{E}(Y|X = x) &= p_1 x + \lambda_2, \\ \mathrm{Var}(Y|X = x) &= x p_1(1 - p_1) + \lambda_2, \end{aligned} \tag{1}$$

where $p_1 = \lambda_{12}/[\lambda_1 + \lambda_{12}]$. For nonnegative integer-valued random variables, the common measure of dispersion is the variance to mean ratio. For (1), the ratio is $1 - x p_1^2/(p_1 x + \lambda_2)$ which is less than 1, and decreases as the dependence in (X, Y) increases (i.e., as p_1 increases).

EXAMPLE 10. Consider a bivariate exponential survival function based on the copula family,

$$C(u, v; \delta) = (u^{-\delta} + v^{-\delta} - 1)^{-1/\delta},$$
$$0 \le u \le 1,\ 0 \le v \le 1, \delta \ge 0.$$

For this family, dependence increases in δ with the independence copula obtaining as $\delta \to 0$ and the Fréchet upper bound obtaining as $\delta \to \infty$. Substitute $u = e^{-x}$ and $v = e^{-y}$ to get

$$\overline{F}(x, y) = C(e^{-x}, e^{-y}; \delta) = (e^{\delta x} + e^{\delta y} - 1)^{-1/\delta}$$

as the survival function for a bivariate exponential random vector (X, Y). The conditional survival function for Y given $X = x$ is

$$\overline{F}_{Y|X}(y|x) = \left[1 - e^{-\delta x} + e^{\delta(y-x)}\right]^{-1-1/\delta}. \tag{2}$$

This is a (zero-)truncated generalized logistic distribution (see Johnson and Kotz, 1970). (2) has scale parameter δ^{-1}, location parameter x and median $\delta^{-1} \log[1 + (2^{\delta/(1+\delta)} - 1)e^{\delta x}]$. Note that with the

Harry Joe is Professor, Department of Statistics, University of British Columbia, Vancouver BC V6T 1Z2 (e-mail: harry@stat.ubc.ca).

positive dependence, the location parameter of the conditional distribution is increasing as x increases, and the dispersion (measured as scale parameter divided by either the median or location parameter) decreases as δ increases.

EXAMPLE 11. The bivariate family with Poisson conditional distributions is a special case of Theorem 3, and given in (42). This family has negative dependence only with correlation range from 0 down to around $-1/3$. For this family, the conditional variance to mean ratio is $\mathrm{Var}(Y|X=x)/\mathrm{E}(Y|X=x) = 1$, so the properties given above also suggest a limited amount of dependence. The univariate margins of this family are overdispersed relative to Poisson, since

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)] \\ &= \mathrm{E}[\mathrm{E}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)] \\ &= \mathrm{E}(Y) + \mathrm{Var}[\mathrm{E}(Y|X)]\end{aligned}$$

so that

$$\frac{\mathrm{Var}(Y)}{\mathrm{E}(Y)} = 1 + \frac{\mathrm{Var}[\mathrm{E}(Y|X)]}{\mathrm{E}(Y)} \geq 1.$$

EXAMPLE 12. The bivariate family with exponential conditional distributions is also a special case of Theorem 3, and given in the equation after (41). This family has negative dependence only with correlation range from 0 down to around -0.3. For this family, the conditional coefficient of variations is $\sqrt{\mathrm{Var}(Y|X=x)}/\mathrm{E}(Y|X=x) = 1$, so the properties given above again suggest a limited amount of dependence. The univariate margins of this family have coefficients of variation larger than 1. With $a(X) = \mathrm{E}(Y|X)$,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}[\mathrm{E}(Y|X)] \\ &= \mathrm{E}[a^2(X)] + \mathrm{Var}[a(X)] \\ &= 2\,\mathrm{Var}[a(X)] + [\mathrm{E}(Y)]^2\end{aligned}$$

so that

$$\frac{\mathrm{Var}(Y)}{[\mathrm{E}(Y)]^2} = 1 + \frac{2\,\mathrm{Var}[a(X)]}{[\mathrm{E}(Y)]^2} \geq 1.$$

EXAMPLE 13. The bivariate family with gamma conditional distributions is also a special case of Theorem 3, and detailed properties are given in Section 4.6 of Arnold et al., 1999. Because gamma distributions have two parameters, this family has a lot more flexibility in the properties of the conditional distributions and regression functions. The dispersions of the conditional distributions have a wide range, there is a wide range of positive and negative correlations, and the regression functions can have a variety of patterns (although if increasing, the regression function reaches an asymptote rather than increase without bound).

As a guideline on other choices of conditional distributions, one could study more conditional distributions of known multivariate families with given margins (cf. Examples 1 and 2), but usually they are not in standard parametric families.

In conclusion, the method of multivariate models based on conditionally specified distributions is quite broad in scope. With this approach, one should pay attention to properties of conditional distributions and how they relate to the type of multivariate dependence that can be attained. Conditional distributions in one-parameter parametric family seem not to provide a flexible class of multivariate models.

Rejoinder

Barry C. Arnold, Enrique Castillo and José María Sarabia

We thank all the discussants for their contributions. Several interesting variations on the conditional specification theme are included in their comments. The additional references that they provide will assist the interested reader in further research in several interesting directions. We will respond to some of their comments though, as will be apparent, there are not any major disagreements to be resolved, except perhaps, in the case of Professor Besag, for questions of pedagogical strategy and style. We begin with some clarifications related to Professor Besag's comments.

Professor Besag begins with an excellent review of conditional specification models with special emphasis on spatial models. Further review of his 1974 paper will certainly reward the reader. It contains many insights and models which have been subsequently utilized and developed by many researchers. That said, it must be emphasized that conditional specification is not inherently a spatial

problem. It is perfectly legitimate and, we believe, pedagogically appropriate to begin with the bivariate case and build up from there to multivariate cases. We are sorry that Professor Besag does not like our topographical map example (presumably that is one of the "contrived" examples that disappointed him). The reader will perhaps be able to envision better introductory examples for use in emphasizing the essential role that conditional densities must play in modeling multivariate phenomena (bivariate, multivariate or even spatial or temporal processes).

Our comment that Professor Besag had discussed distributions with conditionals in exponential families in a specified stochastic process setting is justified by the fact that the title, the first sentence and the general thrust of the 1974 article clearly are focused on spatial processes. The ideas do not have to arise in a stochastic process setting, but Professor Besag's contributions certainly seem to have done so.

The auto-Poisson, autologistic, etc. models discussed in Professor Besag's 1974 paper are intimately related to the multivariate models discussed in our Section 7. Indeed we mention there that they can be viewed as being motivated by certain spatial models. It should be noted that the automodels introduced in the 1974 paper involve a restricted degree of dependence between the variables (related to neighboring sites) and consequently the multivariate models in Section 7 are actually more general than the 1974 automodels (except in the bivariate case where they coincide). We do plead guilty to having selected our examples in Section 6 carefully. If on the other hand Professor Besag's "careful selection" comment was directed not at the choice of examples but at the choice of references given in Section 6, we plead guilty there, too. We referred only to him and to us (and our coauthors). In either case we feel that our careful selection was fair and appropriate. If more examples are desired, reference may be made to Arnold (1987, 1995) and Arnold, Castillo and Sarabia (1993a, b, 1996, 1998).

The practical utility of certain automodels is questioned by Professor Besag. For example, densities with exponential conditionals can only have negative correlations. But this surely complements the perhaps equally restrictive positive correlation usually encountered in marginally specified bivariate exponential densities. As Professor Joe points out, having just one parameter available to model "interaction" does lead to restricted flexibility. However many conditionally specified models actually involve a surfeit of "interaction" parameters.

Perhaps pseudolikelihood is a creature of the 1970s and the 1980s as Professor Besag suggests. However, it does seem to perform well in finite-dimensional settings (a recent paper by Geys, Molenberghs and Ryan, 1999, is quite enthusiastic about its performance in a high-dimensional situation). Recent progress with regard to method of moments estimation for conditionally specified models is described in Arnold, Castillo and Sarabia (2001a), using a multivariate version of Stein's identity.

The extension of the results of Section 4 to multivariate settings that Professor Besag mentions at the end of his comments has been accomplished in a series of recent reports of ours (Arnold, Castillo and Sarabia, 2001b, c, d).

We do apologize to Professor Besag for any errors in our references to his important work in this area; nevertheless, we reserve the right to present the material using what we feel is a pedagogically sound structure even though, clearly, it is not the structure that he would have selected.

Professors Gelman and Raghunathan appear to share Professor Besag's view that spatial applications provide the major arena for conditionally specified models. We feel that their potential in lower-dimensional settings should not be underestimated. Professors Gelman and Raghunathan provide interesting insights into the potential role of conditional specification in missing data imputation. We are initially somewhat disturbed by their enthusiasm for using incompatible conditional specifications. Even though applied collaborators may be happy with possibly inconsistent conditional models, we feel that they should be informed of the degree of incompatibility and perhaps should be persuaded to accept a consistent specification that is in some sense minimally incompatible with the given inconsistent specification. Further investigation of practical problems associated with incompatible specification is clearly called for. Does theory have to catch up with practice here as suggested by Professors Gelman and Raghunathan, or does practice need to catch up to theory? Perhaps both. Until more evidence is in, we confess to a vague distrust of results obtained via inconsistent conditional specifications.

In response to Professor Gelman and Raghunathan's suggestion of a formal Bayesian counterpart to pseudolikelihood, we mention that inconclusive discussion of what is called a pseudo-Bayes (how many different definitions of pseudo-Bayes are there?) approach may be found in Arnold and Press (1991).

Professor Joe provides an informative discussion of dependence phenomena encountered in

conditionally and marginally specified distributions. His comments nicely complement our discussion. He remarks that, as mentioned above, limited correlation flexibility in conditional models is typically encountered when only one "interaction" parameter is present. Postulating that conditional densities belong to multiparameter (rather than one-parameter) exponential families leads to more flexibility in modeling dependence.

Again, we would like to express our gratitude to all the discussants. The ideas, the references and the disagreements that they present should help the reader in further thought, reading, research and application of conditional specification concepts.

ADDITIONAL REFERENCES

ARNOLD, B. C. (1987). Bivariate distribution with Pareto conditionals. *Statist. Probab. Lett.* **5** 263–266.

ARNOLD, B. C. (1995). Conditional survival models. In *Recent Advances in Life-Testing and Reliability, a Volume in Honor of Alonzo Clifford Cohen Jr.* (N. Balakrishnan, ed.) 589–601. CRC Press, Boca Raton, FL.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1993a). Conjugate exponential family priors for exponential family likelihoods. *Statistics* **25** 71–77.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1993b). Multivariate distributions with generalized Pareto conditionals. *Statist. Probab. Lett.* **17** 361–368.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1996). Priors with convenient posteriors. *Statistics* **28** 347–354.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1998). Some alternative bivariate Gumbel models. *Environmetrics* **9** 599–616.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001a). A multivariate version of Stein's identity with applications to moment calculations and estimation of conditionally specified distributions. *Comm. Statist., Theory Methods.* To appear.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001b). Quantification of incompatibility of conditional and marginal information. *Comm. Statist., Theory Methods* **30** 381–395.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001c). Compatibility of partial or complete conditional probability specifications. To appear.

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (2001d). Exact and near compatibility of discrete conditional distributions. To appear.

ARNOLD, B. C. and PRESS, S. J. (1991). Pseudo-Bayesian estimation. Technical Report 186. Dept. Statistics, Univ. California, Riverside.

BADDELEY, A. J. (2000). Time-invariance estimating equations. *Bernoulli* **6** 783–808.

BADDELEY, A. J. (2001). Likelihoods and pseudolikelihoods for Markov spatial processes. In *State of the Art in Probability and Statistics; Festschrift for Willem R. van Zwet* (M. C. M. de Gunst, C. A. J. Klaassen and A. W. van der Vaart, eds.). *IMS Lecture Notes* **36** 21–49.

BADDELEY, A. J. and TURNER, R. (2000). Practical maximum psuedolikelihood for spatial point patterns (with discussion). *Austral. New Zealand J. Statist.* **42** 283–322.

BARTLETT, M. S. (1938). The approximate recovery of information from field experiments with large blocks. *J. Agricultural Science* **28** 418–427.

BARTLETT, M. S. (1967). Inference and stochastic processes. *J. Roy. Statist. Soc. Ser. A* **130** 457–477.

BARTLETT, M. S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 147–174.

BESAG, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

BESAG, J. E. (1975). Statistical analysis of nonlattice data. *The Statistician* **24** 179–195.

BESAG, J. E. (1977). On spatial temporal models and Markov fields. In *Proceedings of the 10th European Meeting of Statisticians, Prague* 47–55. Academia Publishing House of the Czechoslovak Academy of Sciences.

BESAG, J. E. (1978). Some methods of statistical analysis for spatial data (with discussion). *Bull. Internat. Statist. Institute* **47** 77–92.

BESAG, J. E. (1986). On the statistical analysis of dirty pictures (with discussion). *J. Roy. Statist. Soc. Ser. B* **48** 259–302.

BESAG, J. E. and HIGDON, D. H. (1999). Bayesian analysis of agricultural field experirments (with discussion). *J. Roy. Statist. Soc. Ser. B* **61** 691–746.

BESAG, J. and KOOPERBERG, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82** 733–746.

CLIFFORD, P. (1990). Markov random fields in statistics. In *Disorder in Physical Systems* (G. Grimmett and D. J. Welsh, eds.) Clarendon Press, Oxford.

COMETS, F. and JANZURA, M. (1998). A central limit theorem for conditionally centered random fields with an application to Markov fields. *J. Appl. Probab.* **35** 608–621.

FRANK, O. and STRAUSS, D. (1986). Markov graphs. *J. Amer. Statist. Assoc.* **81** 832–842.

GARFINKEL, I. and MEYERS, M. K. (1999). A tale of many cities: the New York City Social Indicators Survey. School of Social Work, Columbia University.

GELMAN, A., KING, G. and LIU, C. (1998). Not asked and not answered: multiple imputation for multiple surveys (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **93** 846–874.

GELMAN, A., VAN MECHELEN, I., VERBECKE, G., HEITJAN, D. F. and MEULDERS, M. (2001). Bayesian model checking for missing and latent data problems using posterior predictive simulations. Technical report, Dept. Statistics, Columbia Univ.

GEYS, H., MOLENBERGHS, G. and RYAN, L. (1999). Pseudolikelihood modeling of multivariate outcomes in developmental toxicology. *J. Amer. Statist. Assoc.* **94** 734–745.

HAMMERSLEY, J. M. and CLIFFORD, P. (1971). Markov fields on finite graphs and lattices. Unpublished manuscript.

HASLETT, J. (1985). Maximum likelihood discriminant analysis on the plane using a Markovian model of spatial content. *Pattern Recognition* **18** 287–296.

HECKERMAN, D., CHICKERING, D. M., MEEK, C., ROUNTHWAITE, R. and KADIE, C. (2000). Dependency networks for inference, collaborative filtering and data visualization. *J. Machine Learning Research* **1** 49–75.

JOE, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.

JOHNSON, N. L. and KOTZ, S. (1970). *Continuous Univariate Distributions* **2** Wiley, New York.

LÉVY, P. (1948). Chaînes doubles de Markoff et functions alétoires de deux variables. *Comptes Rendues de l'Académie des Sciences* **226** 53–55.

LIU, C. (1995). Monotone data augmentation using the multivariate t distribution. *J. Multivariate Anal.* **53** 139–158.

MENG, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statist. Sci.* **9** 538–573.

NELSEN, R. B. (1999). *An Introduction to Copulas*. Springer, New York.

RAGHUNATHAN, T. E. and GRIZZLE, J. E. (1995). A split questionnaire survey design. *J. Amer. Statist. Assoc.* **90** 55–63.

RAGHUNATHAN, T. E., LEPKOWSKI, J. E., SOLENBERGER, P. W. and VAN HOEWYK, J. H. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*. To appear.

RAGHUNATHAN, T. E., SOLENBERGER, P. W. and VAN HOEWYK, J. H. (1997). IVEware: imputation and variance estimation software. Available at `http://www.isr.umich.edu/src/smp/ive`.

RIPLEY, B. D. (1988). *Statistical Inference for Spatial Processes*. Cambridge Univ. Press.

RUBIN, D. B. (1996). Multiple imputation after 18+ years (with discussion). *J. Amer. Statist. Assoc.* **91** 473–520.

SCHAFER, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

TJELMELAND, H. and BESAG, J. E. (1998). Markov random fields with higher-order interactions. *Scand. J. Statist.* **25** 415–433.