This article was downloaded by: [Alzahra University]

On: 10 April 2010

Access details: Access Details: [subscription number 794527431]

Publisher Taylor & Francis

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK

Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713597238>

A Convergent Iterative Procedure for Constructing Bivariate Distributions

Sedigheh Shams ^a

^a Department of Mathematics, Alzahra University, Tehran, Iran

Online publication date: 01 March 2010

To cite this Article Shams, Sedigheh(2010) 'A Convergent Iterative Procedure for Constructing Bivariate Distributions', Communications in Statistics - Theory and Methods, 39: 6, 1026 – 1037

To link to this Article: DOI: 10.1080/03610920902846521

URL: <http://dx.doi.org/10.1080/03610920902846521>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

A Convergent Iterative Procedure for Constructing Bivariate Distributions

SEDIGHEH SHAMS

Department of Mathematics, Alzahra University, Tehran, Iran

For many years there has been interest in families of bivariate distributions with the marginals as parameters. Questions of this kind arise if one is to build a stochastic model in a situation where one has some idea about the dependence structure and marginal distributions. In this article, among all bivariate distributions which satisfy the constraints imposed by the known marginals and/or dependence structure, one that has the maximum entropy is obtained by using iterative procedure, and its convergence is proved.

Keywords Bivariate distribution; Contingency tables; Convergence; Correlation matrix; Iterative procedure; Marginal distribution; Maximum entropy.

Mathematics Subject Classification Primary 62H05; Secondary 62H99.

1. Introduction

For many years there has been interest in families of multivariate distributions with the marginals as parameters. Also, if one has some idea about the kind of dependence, how can you build a stochastic model subject to these information? There are many methods for constructing multivariate distributions subject to information about marginals and dependence structure. In this article, we attempt to construct a bivariate distribution which has the least information about what is not known, or has the maximum entropy subject about what is known.

Shannon (1948), Soofi et al. (1995), Kapour (1989), Zellner and Highfield (1988), Bedford and Meeuwissen (1996), studied various maximum entropy distributions in special classes.

In this article, Sec. 2 describes the concepts of entropy and maximum entropy, and some well-known distributions as maximum entropy distributions in special classes are introduced. In Sec. 3, the approach of finding maximum entropy distribution with given marginals and given correlation matrix is described. Since this approach needs an iterative procedure, this procedure is given and

Received April 15, 2008; Accepted February 24, 2009

Address correspondence to Sedigheh Shams, Department of Mathematics, Alzahra University, Tehran, Iran; E-mail: shams@src.ac.ir

the convergence of it is proved. In Sec. 4, one application of the iterative procedure—obtaining some special contingency tables—is discussed.

2. Entropy and Maximum Entropy Distribution

Shannon (1948) introduced two important ideas of information. The first idea is that information is a statistical concept. The second idea implies that on the basis of the frequency distribution there is an essentially unique function of the distribution which measures the amount of the information. The entropy of a random vector $X = (X_1, \dots, X_n)$ is

$$H(X) = H[f(X)] = - \int \cdots \int f(x_1, \dots, x_n) \log f(x_1, \dots, x_n) dx_1, \dots, dx_n. \quad (1)$$

The Maximum Entropy (ME) model $f^*(x)$ is the density that maximize $H(f)$ subject to the information constraints.

In parametric inference about a vector of unknown parameters $\theta = (\theta_1, \dots, \theta_m)$, Shore and Johnson (1980) considered a class of distributions

$$\Omega_\theta = \{f(x; \theta); E_f[T_j(X)] = \theta_j, j = 1, \dots, m\},$$

where T_j 's are absolutely integrable functions with respect to f . For the continuous case, the inference is based on the model that maximizes the entropy $H[f(X; \theta)] = - \int f(x; \theta) \log f(x; \theta) dx$ subject to the information constraints that define Ω_θ . The ME model $f^*(x; \theta)$ in Ω_θ , if it exists (Soofi et al., 1995), is of the form

$$f^*(x; \theta) = C(\theta) \exp[\eta_1(\theta)T_1(x) + \cdots + \eta_m(\theta)T_m(x)], \quad (2)$$

where $C(\theta)$ is the normalizing constant and $\eta_1, \dots, \eta_m$ are Lagrange multipliers.

Many well-known distributions are ME subject to various types of constraints; see Soofi et al. (1995), Kapour (1989), and Zellner and Highfield (1988).

Suppose Ω_f is the class of all n -variate density functions with fixed marginals,

$$\Omega_f = \{f(x_1, \dots, x_n); f_i(x_i) \ (i = 1, \dots, n) \text{ are its marginals}\}. \quad (3)$$

In order to find maximum entropy distribution subject to Ω_f , define $T_{x_i}(X) = \delta(x_i - X_i)$, ($i = 1, \dots, n$), where δ is the Delta-Dirac function, so $E_f[T_{x_i}(X)] = f_i(x_i)$ for each x_i , ($i = 1, \dots, n$). By these definitions Ω_f can be rewritten as

$$\Omega_f = \{f(x_1, \dots, x_n); E_f[T_{x_i}(X)] = f_i(x_i), i = 1, \dots, n\}. \quad (4)$$

Using Lagrange method for maximizing H subject to Ω_f , the ME density function is the case of independence as follows:

$$f^*(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i), \quad (5)$$

where f_i for $i = 1, \dots, n$, are known marginals.

Now, let $X = (X_1, \dots, X_n)$ be a random vector with given mean vector $\mu = (\mu_1, \dots, \mu_n)$ and covariance matrix Σ . Define $\Omega_{\mu, \Sigma}$ class of all multivariate density functions $f(x_1, \dots, x_n)$ with fixed mean vector μ and covariance matrix Σ ,

$$\Omega_{\mu, \Sigma} = \{f(x_1, \dots, x_n); E_f(X) = \mu, V_f(X) = \Sigma\}. \quad (6)$$

Bedford and Meeuwissen (1996) showed that the ME multivariate distribution subject to $\Omega_{\mu, \Sigma}$ is $MN(\mu, \Sigma)$, multivariate normal distribution with mean vector μ and covariance matrix Σ .

3. ME Multivariate Distribution with Given Marginals and Correlation Matrix

Let $X = (X_1, \dots, X_n)$ be a random vector with given marginal distribution functions $F_1, \dots, F_n$ and given covariance matrix Σ . Define $\Omega_{F, \Sigma}$ the class of all multivariate density functions $f(x_1, \dots, x_n)$ with fixed marginal densities $f_1, \dots, f_n$ and covariance matrix Σ .

The problem is to maximize (1) subject to $\Omega_{F, \Sigma}$. In order to solve this problem, again let $T_{x_i}(X) = \delta(x_i - X_i)$, $(i = 1, \dots, n)$ and $V(X) = (X - \mu)(X - \mu)'$, where $\mu = E(X)$ (μ is known because marginals are known), and V is the covariance matrix of vector X .

Using Lagrange method to maximize H subject to $\Omega_{F, \Sigma}$, the maximizing function is

$$f^*(x_1, \dots, x_n) = \exp \left\{ \sum_{i=1}^n \int \tau_i(u_i) \delta(x_i - u_i) du_i + x'Ax \right\} [M(\tau_1, \dots, \tau_n, A)]^{-1},$$

where $M(\tau_1, \dots, \tau_n, A) = \int \dots \int \exp \left\{ \sum_{i=1}^n \tau_i(x_i) + x'Ax \right\} dx_1, \dots, dx_n$. By setting $\exp\{\tau_i(x_i)\} = d_i(x_i)$, $(i = 1, \dots, n)$ and normalizing such that $M(\tau_1, \dots, \tau_n, A) = 1$,

Table 1
Examples of maximum entropy distributions

Support	T(X)	ME distribution
(a, b)	None	Uniform
$(0, 1)$	$\log(X), \log(1 - X)$	Beta
$(0, \infty)$	X	Exponential
$(0, \infty)$	$X, \log X$	Gamma
$(0, \infty)$	$X^\beta, \log X, (\beta \neq 1)$	Weibull
$(a, \infty), a > 0$	$\log X$	Pareto
$(-\infty, \infty)$	$ X $	Laplace
$(-\infty, \infty)$	X^2	Normal (mean = 0)
$(-\infty, \infty)$	X, X^2	Normal
$(-\infty, \infty)$	$\log(1 + X^2)$	Generalized Cauchy
$(-\infty, \infty)$	$X, \log(1 + e^{-X})$	Generalized Logistic
$(-\infty, \infty)$	X, e^{-X}	Generalized Extreme value
$(-\infty, \infty)$	X, X^2, X^3, X^4	Quadratic Exponential

the maximizing density is

$$f^*(x_1, \dots, x_n) = \prod_{i=1}^n d_i(x_i) \cdot \exp\{x'Ax\}, \quad (7)$$

where d_i , ($i = 1, \dots, n$) and A are such that

$$\begin{aligned} f_i(x_i) &= d_i(x_i) \int \cdots \int \prod_{j \neq i} d_j(x_j) \exp\{x'Ax\} dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_n \\ \Sigma &= \int \cdots \int (x - \mu)(x - \mu)' \prod_{i=1}^n d_i(x_i) \exp\{x'Ax\}. \end{aligned} \quad (8)$$

In a special case when Σ is a diagonal matrix, named Σ^* , it is shown that the ME distribution subject to Ω_{F, Σ^*} is the independent case.

Theorem 3.1. *The ME multivariate distribution that maximizes H subject to Ω_{F, Σ^*} , is the case of independence.*

Proof. Suppose $f^*(x_1, \dots, x_n) = \prod_{i=1}^n f_i(x_i)$. It is obvious that $f^* \in \Omega_{F, \Sigma^*}$ and is ME subject to Ω_F . Since $f^* \in \Omega_F$ and $\Omega_{F, \Sigma^*} \subseteq \Omega_F$ hence f^* is the ME density subject to Ω_{F, Σ^*} . $\square$

Theorem 3.2. *The density that maximizes H subject to $\Omega_{F, \Sigma}$, is of the form (7), (8).*

Proof. Let $f = f^*$ and g be any density function belongs to $\Omega_{F, \Sigma}$, then $h_\lambda = (1 - \lambda)f + \lambda g$, ($0 \leq \lambda \leq 1$) belongs to $\Omega_{F, \Sigma}$, too. The entropy of h_λ is

$$H[h_\lambda] = - \int \cdots \int h_\lambda(x_1, \dots, x_n) \log h_\lambda(x_1, \dots, x_n) dx_1, \dots, dx_n$$

It follows that the derivative of the entropy with respect to λ equals

$$\frac{\partial}{\partial \lambda} H[h_\lambda] = \int \cdots \int [f(x_1, \dots, x_n) - g(x_1, \dots, x_n)] \log h_\lambda(x_1, \dots, x_n) dx_1, \dots, dx_n.$$

Since $\log f = \sum \log b_i(x_i) + x'Ax$ and f, g have the same marginals and the same covariance matrix at $\lambda = 0$ the derivative equals to zero:

$$\int \cdots \int [f(x_1, \dots, x_n) - g(x_1, \dots, x_n)] \log f(x_1, \dots, x_n) dx_1, \dots, dx_n = 0.$$

Hence, $H[h_\lambda]$ has a stationary point at $\lambda = 0$, and this stationary point is a maximum because

$$\frac{\partial^2}{\partial \lambda^2} H[h_\lambda] = - \int \cdots \int \frac{[f(x_1, \dots, x_n) - g(x_1, \dots, x_n)]^2}{h_\lambda(x_1, \dots, x_n)} dx_1, \dots, dx_n < 0.$$

Thus, for all $g \in \Omega_{F, \Sigma}$, $H[f^*] \leq H[g]$. It will suffice to exhibit a density of the form (7) and having marginals and Σ to have ME. $\square$

To find $d_i(x_i)$, $(i = 1, \dots, n)$ and A such that (8) holds, it is required an iterative procedure. The following is the formulation of an iterative procedure which will be shown that it converges. For proving convergence of the procedure we use a well-known information-theoretic measure of discrepancy between distributions.

The Kullback discrimination information function between two distributions is

$$K[f; g] = \int f(x) \log \frac{f(x)}{g(x)} dx = -H[f] - E_f[\log(g(x))]. \quad (9)$$

It is well known that $K[f; g] \geq 0$ and the equality holds if and only if $f(x) = g(x)$ almost everywhere.

3.1. Iterative Procedure

Let $n = 2$; the ME density subject to $\Omega_{F, \Sigma}$ is

$$f^*(x_1, x_2) = d_1(x_1)d_2(x_2)e^{a_1x_1^2+a_{x_1x_2}+a_2x_2^2} \quad (10)$$

since marginals and $\text{cov}(X, Y)$ are known, (8) reduces to

$$\begin{aligned} f_1(x_1) &= b_1(x_1) \int b_2(x_2) e^{ax_1x_2} dx_2 \\ f_2(x_2) &= b_2(x_2) \int b_1(x_1) e^{ax_1x_2} dx_1 \\ c &= \iint x_1x_2 b_1(x_1)b_2(x_2) e^{ax_1x_2} dx_1 dx_2, \end{aligned}$$

where $b_i(x_i) = e^{a_ix_i^2} d_i(x_i)$ and the iteration is given by

$$\begin{cases} f^{(2k-1)}(x_1, x_2) = f_1(x_1)[f_1^{(2k-2)}(x_1)]^{-1} f^{(2k-2)}(x_1, x_2) \\ f^{(2k)}(x_1, x_2) = f_2(x_2)[f_2^{(2k-1)}(x_2)]^{-1} f^{(2k-1)}(x_1, x_2) \\ f^{(0)}(x_1, x_2) = b_1^{(0)}(x_1)b_2^{(1)}(x_2)e^{ax_1x_2} \end{cases} \quad (11)$$

By choosing $b_1^{(0)}(x_1)$ and $b_2^{(1)}(x_2)$ as initial function in (11) and assuming

$$\begin{cases} b_1^{(i)}(x_1) = \frac{f_1(x_1)}{\int b_2^{(i)}(x_2) e^{ax_1x_2} dx_2} \\ b_2^{(j)}(x_2) = \frac{f_2(x_2)}{\int b_1^{(j-1)}(x_1) e^{ax_1x_2} dx_1} \end{cases},$$

it is seen that

$$\begin{cases} f_1^{(2k-1)}(x_1) = f_1(x_1), \quad f_2^{(2k)}(x_2) = f_2(x_2), \\ \begin{cases} f^{(2k-1)}(x_1, x_2) = b_1^{(k)}(x_1)b_2^{(k)}(x_2)e^{ax_1x_2} \\ f^{(2k)}(x_1, x_2) = b_1^{(k)}(x_1)b_2^{(k+1)}(x_2)e^{ax_1x_2} \end{cases} \end{cases} \quad (12)$$

so that the iterated functions are of the form (7).

Theorem 3.3. *Let (10) and (12) hold; then*

$$f^{(N)}(x_1, x_2) = f^*(x_1, x_2) \quad \text{a.e.}$$

Proof. Consider

$$K[f^*; f^{(2k)}] = K[f^*; f^{(2k-1)}] - K[f_2; f_2^{(2k-1)}] \quad (13)$$

and

$$K[f^*; f^{(2k+1)}] = K[f^*; f^{(2k)}] - K[f_1; f_1^{(2k)}]. \quad (14)$$

Since the discrimination information values in (13) and (14) are non negative, $K[f^*; f^{(2k)}] \leq K[f^*; f^{(2k-1)}]$, with equality holds if and only if

$$f_2(x_2) = f_2^{(2k-1)}(x_2) \quad \text{a.e.}, \quad (15)$$

and $k[f^*; f^{(2k+1)}] \leq K[f^*; f^{(2k)}]$, with equality holds if and only if

$$f_1(x_1) = f_1^{(2k)}(x_1) \quad \text{a.e.} \quad (16)$$

It is clear that

$$K[f^*; f^{(1)}] \geq K[f^*; f^{(2)}] \geq \dots \geq K[f^*; f^{(2k-1)}] \geq K[f^*; f^{(2k)}] \geq \dots \geq 0. \quad (17)$$

Let us consider when there is equality in someplace in (17), say,

$$K[f^*; f^{(2k-1)}] = K[f^*; f^{(2k)}];$$

then from (15), (16), and the first row of (12) it follows that *almost everywhere*

$$f_1^{(2k)}(x_1) = f_1^{(2k-1)}(x_1) = f_1(x_1), \quad f_2^{(2k+1)}(x_2) = f_2^{(2k)}(x_2) = f_2(x_2) \quad (18)$$

and that is equality thereafter, hence for $N \geq 2k - 1$, *almost everywhere*

$$f_1^{(N)}(x_1) = f_1(x_1), \quad f_2^{(N)}(x_2) = f_2(x_2), \quad f^{(N)}(x_1, x_2) = b_1^{(N)}(x_1)b_2^{(N)}(x_2)e^a x_1 x_2. \quad (19)$$

It will now be shown that $f^{(N)}(x_1, x_2) = f^*(x_1, x_2)$, a.e., ($N \geq 2k - 1$). Since $f^*(x_1, x_2)$ minimizes $\iint f^*(x_1, x_2) \log f^*(x_1, x_2)$ for all densities with marginals $f_1(x_1)$, $f_2(x_2)$ and fixed covariance,

$$\begin{aligned} & \iint f^{(N)}(x_1, x_2) \log f^{(N)}(x_1, x_2) e^{-ax_1 x_2} dx_1 dx_2 \\ & \geq \iint f^*(x_1, x_2) \log f^*(x_1, x_2) e^{-ax_1 x_2} dx_1 dx_2 \\ & = \iint f^*(x_1, x_2) \log f^*(x_1, x_2) e^{-ax_1 x_2} dx_1 dx_2 \\ & \quad + \iint f^*(x_1, x_2) \log f^{(N)}(x_1, x_2) e^{-ax_1 x_2} dx_1 dx_2 \end{aligned}$$

but $\log f^{(N)}(x_1, x_2)e^{-x_1x_2} = \log b_1^{(N)}(x_1) + \log b_2^{(N)}(x_2)$, so that

$$\begin{aligned} & \iint f^{(N)}(x_1, x_2) \log f^{(N)}(x_1, x_2) e^{-ax_1x_2} dx_1 dx_2 \\ &= \int f_1(x_1) \log b_1^{(n)}(x_1) dx_1 + \int f_2(x_2) \log b_2^{(N)}(x_2) dx_2 \\ &= \iint f^*(x_1, x_2) \log f^{(N)}(x_1, x_2) e^{-ax_1x_2} dx_1 dx_2 \end{aligned}$$

and it is clear that

$$0 \geq \iint f^*(x_1, x_2) \log \frac{f^*(x_1, x_2)}{f^{(N)}(x_1, x_2)} dx_1 dx_2 = K[f^*; f^{(N)}]. \quad (20)$$

Since $K[f^*; f^{(N)}]$ should be nonnegative, hence

$$K[f^*, f^{(N)}] = 0$$

and

$$f^*(x_1, x_2) = f^{(N)}(x_1, x_2) \text{ a.e.} \quad (21)$$

Now consider the case when there is no equality in (17). Since (17) is a monotone decreasing sequence of non negative numbers bounded below, it converges to a finite value as $k \rightarrow \infty$, hence from (13) and (14)

$$K[f^*; f^{(2k-1)}] - K[f^*; f^{(2k)}] = K[f_2; f_2^{(2k-1)}] \rightarrow 0 \quad (22)$$

$$K[f^*; f^{(2k)}] - K[f^*; f^{(2k+1)}] = K[f_1; f_1^{(2k)}] \rightarrow 0. \quad (23)$$

Kullback (1959) showed that as $N \rightarrow \infty$ (22), (23) imply

$$\int |f_1^{(N)}(x_1) - f_1(x_1)| dx_1 \rightarrow 0, \quad \int |f_2^{(N)}(x_2) - f_2(x_2)| dx_2 \rightarrow 0. \quad (24)$$

Hence, using (12) as $k \rightarrow \infty$,

$$\begin{aligned} & \iint |f^{(2k)}(x_1, x_2) - f^{(2k-1)}(x_1, x_2)| dx_1 dx_2 \\ &= \iint f^{(2k-1)}(x_1, x_2) |f_2(x_2) - f_2^{(2k-1)}(x_2)| [f_2^{(2k-1)}(x_2)]^{-1} dx_1 dx_2 \\ &= \int |f_2(x_2) - f_2^{(2k-1)}(x_2)| dx_2 \rightarrow 0 \end{aligned} \quad (25)$$

and

$$\begin{aligned} & \iint |f^{(2k+1)}(x_1, x_2) - f^{(2k)}(x_1, x_2)| dx_1 dx_2 \\ &= \iint f^{(2k)}(x_1, x_2) |f_1(x_1) - f_1^{(2k)}(x_1)| [f_1^{(2k)}(x_1)]^{-1} dx_1 dx_2 \\ &= \int |f_1(x_1) - f_1^{(2k)}(x_1)| dx_1 \rightarrow 0 \end{aligned} \quad (26)$$

and for any m , as $N \rightarrow \infty$

$$\begin{aligned} & \iint |f^{(N+m)}(x_1, x_2) - f^{(N)}(x_1, x_2)| dx_1 dx_2 \\ & \leq \iint |f^{(N+m)}(x_1, x_2) - f^{(N+m-1)}(x_1, x_2)| dx_1 dx_2 \\ & \quad + \iint |f^{(N+m-1)}(x_1, x_2) - f^{(N+m-2)}(x_1, x_2)| dx_1 dx_2 \\ & \quad + \cdots + \iint |f^{(N+1)}(x_1, x_2) - f^{(N)}(x_1, x_2)| dx_1 dx_2 \rightarrow 0, \end{aligned} \quad (27)$$

so that there exists a function denoted by $f^{(\infty)}(x_1, x_2)$ defined uniquely a.e. (see Titchmarsh, 1939), such that

$$\iint |f^{(N)}(x_1, x_2) - f^{(\infty)}(x_1, x_2)| dx_1 dx_2 \rightarrow 0, \quad (28)$$

also

$$\int |f_1^{(N)}(x_1) - f_1^{(\infty)}(x_1)| dx_1 \rightarrow 0, \quad \int |f_2^{(N)}(x_2) - f_2^{(\infty)}(x_2)| dx_2 \rightarrow 0 \quad (29)$$

as $N \rightarrow \infty$,

$$\int |f_1(x_1) - f_1^{(\infty)}(x_1)| dx_1 \leq \int |f_1(x_1) - f_1^{(N)}(x_1)| dx_1 + \int |f_1^{(N)}(x_1) - f_1^{(\infty)}(x_1)| dx_1 \rightarrow 0 \quad (30)$$

and

$$\int |f_2(x_2) - f_2^{(\infty)}(x_2)| dx_2 \leq \int |f_2(x_2) - f_2^{(N)}(x_2)| dx_2 + \int |f_2^{(N)}(x_2) - f_2^{(\infty)}(x_2)| dx_2 \rightarrow 0. \quad (31)$$

It follows that

$$\begin{aligned} & \int |f_1(x_1) - f_1^{(\infty)}(x_1)| dx_1 = 0, \quad f_1(x_1) = f_1^{(\infty)}(x_1), \quad \text{a.e.}, \\ & \int |f_2(x_2) - f_2^{(\infty)}(x_2)| dx_2 = 0, \quad f_2(x_2) = f_2^{(\infty)}(x_2), \quad \text{a.e.} \end{aligned} \quad (32)$$

It can be shown as before that $f^{(\infty)}(x_1, x_2) = f^*(x_1, x_2)$ almost everywhere. $\square$

Example 3.1. Let X_1 and X_2 be two random variables with standard normal distributions and correlation ρ . Using iteration (12) and assuming $b_1^{(0)}(x_1) = e^{-x_1^2/2}$ and $b_2^{(1)}(x_2) = e^{-x_2^2/2}$ we will have

$$\begin{aligned} b_1^{(1)}(x_1) &= \frac{1}{\sqrt{(2\pi)}} e^{-(1+a^2)x_1^2/2}, \\ b_2^{(1)}(x_2) &= \frac{1}{\sqrt{(2\pi)}} e^{-(1+a^2)x_2^2/2} : \\ f_{(x_1, x_2)}^{(1)} &= \frac{1}{2\pi} e^{-(1+a^2)\left(\frac{x_1^2}{2} + \frac{x_2^2}{2}\right) + ax_1x_2}; \end{aligned}$$

in a second step,

$$\begin{aligned} b_1^{(2)}(x_1) &= \frac{\sqrt{(1+a^2)}}{\sqrt{(2\pi)}} e^{-(1+\frac{a^2}{2+2a^2})x_1^2/2}, \\ b_2^{(2)}(x_2) &= \frac{\sqrt{(1+a^2)}}{\sqrt{(2\pi)}} e^{-(1+\frac{a^2}{2+2a^2})x_2^2/2}, \\ f_{(x_1, x_2)}^{(2)} &= \frac{\sqrt{(1+a^2)}}{2\pi} e^{-(1+a^2)\frac{x_1^2}{2} - (1+a^2/(2+2a^2))\frac{x_2^2}{2} + ax_1x_2}, \end{aligned}$$

in a third step,

$$\begin{aligned} b_1^{(3)}(x_1) &= \frac{\sqrt{(1+a^2/(2+2a^2))}}{\sqrt{(2\pi)}} e^{-(1+\frac{a^2}{1+a^2/(2+2a^2)})x_1^2/2}, \\ b_2^{(3)}(x_2) &= \frac{\sqrt{(1+a^2/(2+2a^2))}}{\sqrt{(2\pi)}} e^{-(1+\frac{a^2}{1+a^2/(2+2a^2)})x_2^2/2}, \\ f_{(x_1, x_2)}^{(3)} &= \frac{1+a^2/(2+2a^2)}{2\pi} e^{-(1+\frac{a^2}{1+a^2/(2+2a^2)})x_1^2/2 - (1+\frac{a^2}{1+a^2/(2+2a^2)})x_2^2/2 + ax_1x_2}. \end{aligned}$$

Since there is another constraint related to knowing ρ , each of the following equations which has proper solution could be the ME bivariate distribution:

$$\begin{aligned} \iint x_1 x_2 f_{(x_1, x_2)}^{(1)} dx_1 dx_2 &= \frac{a}{(\sqrt{(1+a^2)})(4+7a^2+4a^4)(\sqrt{(4+7a^2+4a^4)/(1+a^2)})} = \rho, \\ \iint x_1 x_2 f_{(x_1, x_2)}^{(2)} dx_1 dx_2 &= \frac{a}{(4+5a^2)(\sqrt{(4+5a^2)/1+a^2})} = \rho \end{aligned}$$

or

$$\iint x_1 x_2 f_{(x_1, x_2)}^{(i)} dx_1 dx_2 = \rho.$$

3.2. Extension to n -Variate

In a general case, for solving (12) the iteration is given by

$$\begin{cases} f^{(nk+1)}(x_1, \dots, x_n) = f_1(x_1)[f_1^{(nk)}(x_1)]^{-1} f^{(nk)}(x_1, \dots, x_n) \\ f^{(nk+2)}(x_1, \dots, x_n) = f_2(x_2)[f_2^{(nk+1)}(x_2)]^{-1} f^{(nk+1)}(x_1, \dots, x_n) \\ \dots \\ f^{(nk+n)}(x_1, \dots, x_n) = f_n(x_n)[f_n^{(nk+n-1)}(x_n)]^{-1} f^{(nk+n-1)}(x_1, \dots, x_n) \\ f^{(0)}(x_1, \dots, x_n) = \prod_{i=1}^n b_i^{(0)}(x_i) e^{x'Ax} \end{cases}. \quad (33)$$

As in bivariate case for any fixed A it can be shown that the iterative procedure converges.

4. Application in Contingency Tables

Applications of contingency tables have had a long history in statistical theory. In a two-dimensional table of I rows and J columns, each entry in the matrix represents an observed frequency. When the row and column totals and the dependence are specified, then only a subset of frequencies can be assigned arbitrary values; the rest will be determined by the given row and column totals, and the dependence. Here, the estimates of frequencies obtained from an application of the maximum entropy are considered.

Let P_{ij} be the probability in the i th row and the j th column of the $I \times J$ contingency table in which the row probability sums are $P_1, \dots, P_I$ and the column probability sums are $Q_1, \dots, Q_J$ and the covariance is C , so that

$$\begin{aligned} P_i &= \sum_{j=1}^J P_{ij}; \quad i = 1, \dots, I, \\ Q_j &= \sum_{i=1}^I P_{ij}; \quad j = 1, \dots, J, \\ C &= \sum_{i=1}^I \sum_{j=1}^J ijP_{ij} - \sum_{i=1}^I iP_i \sum_{j=1}^J jQ_j. \end{aligned} \quad (34)$$

Suppose the values of the row probability sums and the column probability sums are known, but the cell probabilities are unknown. Since the sum of row totals is equal to the sum of column totals, there are $I + J$ independent constraints on the cell probabilities. However, there are IJ unknown probabilities and only $I + J$ constraints. So $IJ - (I + J)$ cell probabilities can be filled arbitrarily, or there are $IJ - (I + J)$ degrees of freedom.

Thus, there can be an infinite set of probabilities that will be consistent with the given information. One of these is the set of probabilities that maximize the entropy

$$H[P] = - \sum_{i=1}^I \sum_{j=1}^J P_{ij} \log P_{ij} \quad (35)$$

subject to constraints (34).

Maximizing H is equivalent to minimizing $-H$, subject to some linear constraints. We can now specialize our discussion to the important class of nonlinear problems termed *convex-programming*.

The Lagrangian function is

$$\begin{aligned} L &= - \sum_{i=1}^I \sum_{j=1}^J P_{ij} \log P_{ij} - \sum_{i=1}^I \lambda_i \left(\sum_{j=1}^J P_{ij} - P_i \right) - \sum_{j=1}^J \gamma_j \left(\sum_{i=1}^I P_{ij} - Q_j \right) \\ &\quad - v \left(\sum_{i=1}^I \sum_{j=1}^J ijP_{ij} - \sum_{i=1}^I iP_i \sum_{j=1}^J jQ_j - C \right) \end{aligned}$$

This gives

$$P_{ij} = A_i B_j P_i Q_j e^{-ijv}, \quad (36)$$

where $A_i = e^{-\lambda_i}$; ($i = 1, \dots, I$), $B_j = e^{-\gamma_j}$, ($j = 1, \dots, J$), and v must be determined such that (34) hold.

The maximum entropy is given by

$$H_{\max} = - \sum_{i=1}^I [P_i \log A_i + P_i \log P_i] + \sum_{j=1}^J [P_j \log B_j + P_j \log P_j] - vC. \quad (37)$$

The Kuhn–Tucker conditions (see Martos, 1975) to find the saddle point are:

$$\begin{aligned} P_{ij} &= A_i B_j P_i Q_j e^{-ijv} \\ 1 &= A_i \sum_{j=1}^J B_j Q_j e^{-ijv}; \quad i = 1, \dots, I \\ 1 &= B_j \sum_{i=1}^I A_i P_i e^{-ijv}; \quad j = 1, \dots, J \\ C &= \sum_{i=1}^I \sum_{j=1}^J ij A_i B_j P_i Q_j e^{-ijv} - \sum_{i=1}^I i P_i \sum_{j=1}^J j Q_j \\ P_{ij} &\geq 0; \quad (i = 1, \dots, I), \quad (j = 1, \dots, J). \end{aligned} \quad (38)$$

As is shown in Lagrange function, v , A_i , $\{i = 1, \dots, I\}$ and B_j , $\{j = 1, \dots, J\}$ are the Lagrangian multipliers, and must be determined. One method for determining them is proposed as follows.

First, take a fixed value of v and solve the second and third equations in (38) using the following iterative scheme to find A_i , $\{i = 1, \dots, I\}$ and B_j , $\{j = 1, \dots, J\}$ (by assuming $A_i^{(0)}$ or $B_j^{(0)}$ arbitrary values, and using this iterative procedure until the values converges):

$$\begin{cases} A_i^{(k+1)} = \left[\sum_{j=1}^J B_j^{(k)} Q_j e^{-ijv} \right]^{-1} \\ B_j^{(k+1)} = \left[\sum_{i=1}^I A_i^{(k)} P_i e^{-ijv} \right]^{-1} \end{cases}. \quad (39)$$

Then by using the forth equation in (38) the value of C would be determined, thus for each value of v , there will be a unique value of C . A table of values of C for corresponding values of v can be prepared, and from this table the appropriate value of v for a given C can be chosen.

Example 4.1. In a contingency table with three rows and four columns, suppose the row and column probabilities are $P_1 = 0.1$, $P_2 = 0.2$, $P_3 = 0.7$; $Q_1 = 0.2$, $Q_2 = 0.3$, $Q_3 = 0.1$, $Q_4 = 0.4$. By using the mentioned approach for the given value of C (after choosing an appropriate value for v) the ME probabilities subject to P_i 's, Q_j 's (here $C = \sum \sum ij P_{ij}$) are showed in Table 2.

Table 2
Values of v , c and ME probabilities subject to marginals and c

	$i \setminus j$	1	2	3	4
$C = 7.59356$ ($v = -4.0$)	1	0.098159	0.001184	0.000000	0.000000
	2	0.098309	0.100698	0.000925	0.000068
	3	0.003530	0.197460	0.099074	0.399931
$C = 7.54923$ ($v = -2.0$)	1	0.086898	0.012975	0.000011	0.000009
	2	0.900852	0.099391	0.006681	0.003843
	3	0.023015	0.187633	0.093200	0.396147
$C = 7.02000$ ($v = 0.0$)	1	0.020000	0.030000	0.010000	0.040000
	2	0.040000	0.060000	0.020000	0.080000
	3	0.140000	0.210000	0.070000	0.280000
$C = 6.53562$ ($v = 2.0$)	1	0.000000	0.000077	0.001198	0.098723
	2	0.000694	0.007523	0.015784	0.175998
	3	0.199305	0.292398	0.083017	0.125277
$C = 6.50380$ ($v = 4.0$)	1	0.000000	0.000000	0.000031	0.099969
	2	0.000002	0.000190	0.003356	0.196452
	3	0.199997	0.299809	0.096612	0.103578

Acknowledgment

This research has been partially supported by Alzahra University.

References

- Bedford, T., Meeuwissen, A. M. H. (1996). The determination of maximum entropy distributions given fixed rank correlation, and numerical approximations for use in sensitivity analysis. *Proceedings SAM095; Theory and Applications of Sensitivity Analysis of Model Output in Computer Simulation*. Office for Official Publications of European Communities, Luxembourg, 19–24.
- Kapour, J. N. (1989). *Maximum Entropy in Science and Engineering*. New York: John Wiley.
- Kullback, S. (1959). *Information Theory and Statistics*. New York: John Wiley.
- Martos, B. (1975). *Nonlinear Programming*. Amsterdam: North-Holland.
- Shannon, C. E. (1948). A mathematical theory of communications. *Tech. J.* 27:370–423, 623–656.
- Shore, J. E., Johnson, R. W. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of the minimum cross-entropy. *IEEE Trans. Inform. Theor.* It-26:26–37.
- Soofi, E. S., Ebrahimi, N., Habibollah, M. (1995). Information distribuishability with application to analysis of failure data. *J. Amer. Statist. Assoc.* 90(430):657–668.
- Titchmarsh, E. C. (1939). *The Theory of Functions*. 2nd ed. Oxford: Oxford University Press.
- Zellner, A., Highfield, R. A. (1988). Calculation of maximum entropy distributions and approximation of marginal posterior distributions. *J. Econometrics* 37:195–209.