# On Joint and Conditional Entropies

**D.V. Gokhale**

Department of Statistics, University of California, Riverside, CA 92521, USA
Tel. (909) 787-3775, Fax (909)787-3286 (Attn. Prof. Gokhale), E-mail: dvgres@ucrac1.ucr.edu

**Abstract:** It is shown that if the conditional densities of a bivariate random variable have maximum entropies, subject to certain constraints, then the bivariate density also maximizes entropy, subject to appropriate constraints. Some examples are discussed.

**Keywords:** Joint and conditional entropies. Maximum entropy subject to constraints.

## Introduction and Notation

Let $f(x,y)$ denote a continuous bivariate probability density defined on the support $S_X \times S_Y$. The entropy of $f(x,y)$ is defined as $H(f) = E_f(-\ln f(X,Y))$. We shall use a similar notation for the entropy of univariate densities. Let the conditional densities of $f(x,y)$ be denoted by $f_1(x|y)$ and $f_2(y|x)$. Many families of probability distributions are known to maximize the entropy among distributions that satisfy given constraints on the expectations of certain statistics. The Beta, Gamma and normal families of distributions are well known examples of this principle. Suppose now that $f_1(x|y)$ and $f_2(y|x)$ are known to belong to such families. The purpose of the present note is to show that $f(x,y)$ also maximizes the entropy subject to constraints on the expectations of suitably defined statistics. This statement is made more specific and more precise in the following section and presented as a theorem.

**Main Results**

Let us make the following assumptions on $f_1(x|y)$ and $f_2(y|x)$.

Assumption 2.1. Density $f_1(x|y)$ maximizes the entropy $H(f_1(y))$, for each $y \, \varepsilon \, S_Y$, in the class $\Psi_1$ of continuous densities which satisfy

$$E_1(q_i(X)) = \mu_i(y), \, i = 1,2,\ldots, k, \tag{1}$$

for some constants $\mu_i$, which may depend on y. Here $E_1$ denotes the expectation with respect to the densities in $\Psi_1$. The statistics $\{q_i(x)\}$ are functionally independent.

Assumption 2.2. Density $f_2(y|x)$ maximizes the entropy $H(f_2(x))$, for each $x \, \varepsilon \, S_X$, in the class $\Psi_2$ of continuous densities which satisfy

$$E_2(t_j(Y)) = \nu_j(x), \, j = 1,2,\ldots, l, \tag{2}$$

for some constants $\nu_j$ , which may depend on x. Here $E_2$ denotes the expectation with respect to the densities in $\Psi_2$. The statistics $\{t_j(y)\}$ are functionally independent.

Theorem 1: If Assumptions 2.1 and 2.2 hold, then the joint density $f(x,y)$ maximizes its entropy $H(f)$ in the class $\Psi$ of densities subject to the constraints

$$E_g[q_i(X))t_j(Y)] = \sigma_{ij}, \, i = 0,1,2,\ldots, k, \, j = 0,1,2,\ldots,l, \tag{3}$$

where $q_0(x) = t_0(y) \equiv 1$, $\sigma_{ij}$ are appropriate constants, and $E_g$ denotes the expectation with respect to densities in $\Psi$.

Proof: It is well known that (see [2,3], for example) under Assumption 2.1 the density $f_1(x|y)$ is of the form

$$f_1(x|y) = \beta_1(\underline{\theta}(y))\exp\{\textstyle\sum_{i=1}^{k} \theta_i(y)q_i(x)\}, \, x \, \varepsilon \, S_X, \, y \, \varepsilon \, S_Y, \tag{4}$$

and equals zero otherwise. The $\mu_i(y)$ and $(\beta_1(\underline{\theta}(y)), \theta_i(y))$ are assumed to be 1:1 functions. Similarly, under Assumption 2.2 the density $f_2(y|x)$ is of the form

$$f_2(y|x) = \beta_2(\underline{\lambda}(x))\exp\{\textstyle\sum_{j=1}^{l} \lambda_j(x)t_j(y)\}, \, y \, \varepsilon \, S_Y, \, x \, \varepsilon \, S_X, \tag{5}$$

and equals zero otherwise. The $\nu_j(x)$ and $(\beta_2(\underline{\lambda}(x)), \lambda_j(x))$ are assumed to be 1:1. Now, if $f_1(x|y)$ and $f_2(y|x)$ are given to be of the forms (4) and (5) respectively, it has been shown by Arnold, et.al, [1, page 36] that the bivariate density $f(x,y)$ must be of the form

$$f(x,y) = \exp\{q(x)'Mt(y)\}, \, (x,y) \, \varepsilon \, S_X \times S_Y, \tag{6}$$

and equals zero otherwise, for a suitable choice of the $(k+1)\times(l+1)$ matrix of parameters M. Here $q(x) = (q_0(x), \, q_1(x),q_2(x),\ldots,q_k(x))'$ and $t(y) = (t_0(y),t_1(y),t_2(y),\ldots,t_l(y))'$. From (6) it follows that $f(x,y)$ maximizes the entropy $H(f)$ in class of continuous bivariate densities $\Psi$, in which, for $g \, \varepsilon \, \Psi$,

$$E_g[q_i(X))t_j(Y)] = \sigma_{ij}, \, i = 0,1,2,\ldots,k, \, j = 0,1,2,\ldots,l, \tag{7}$$

and the parameters in M are 1:1 functions of $\sigma_{ij}$. Note that the maximum entropy density (6) incorporates constraints on the mixed moments of the form $E_g[q_i(X))t_j(Y)]$ only.

## Examples and some remarks

As first example, suppose that the conditional density of Y given X = x is normal with mean a(x) and standard deviation b(x), and the conditional density of X given Y = y is normal with mean d(y) and standard deviation c(y). Arnold <u>et.al.</u> [1, pages 25-27] show that the joint density of (X,Y) must be of the form

$$f(x,y) = (2\pi)^{-1}\exp\{-[A+2Bx+2Gy+Cx^2+Dy^2+2Hxy+2Jx^2y+2Exy^2+Fx^2y^2]/2\}, \tag{8}$$

in which the parameters A through F must satisfy one of the two sets of the following conditions.

(i)   $F = E = J = 0$, $D > 0$, $C > 0$, $H^2 < DC$, or

(ii)   $F > 0$, $FD > E^2$, $CF > J^2$.

Models satisfying conditions (i), are classical bivariate normal densities with normal marginals, normal conditionals, linear regressions and constant conditional variances. Models satisfying conditions (ii) have normal conditionals (as specified) but distinctly non-normal densities. They are

$$g(x) = [2\pi(D+2Ex+Fx^2)]^{-1/2}\exp\{-[A+2Bx+Cx^2 - (G+Hx+Jx^2)^2/(D+2Ex+Fx^2)]/2\}$$

for X and

$$h(y) = [2\pi(C+2Jy+Fy^2)]^{-1/2}\exp\{-[A+2Gy+Dy^2 - (B+Hy+Ey^2)^2/(C+2Jy+Fy^2)]/2\}$$

for Y. From the point of view of this paper, these results provide an interesting insight in the structure of joint maximum entropy distributions when conditional maximum entropy distributions are specified. Note that one set of conditions give the usual bivariate family of normal distributions, which is well known to maximize entropy subject to constraints on the moments given by a given mean vector and a positive definite covariance matrix. In such families the joint moment constraint is through the covariance of X and Y while the other constraints are on the marginal moments. The other set of conditions yields a family of bivariate distributions which still has maximum entropy, but the marginal moment constraints are remarkably different and complicated.

As second example, consider the situation when the conditional densities of X|y and Y|x are exponential with mean functions given by $[\mu(y)]^{-1}$ and $[\nu(x)]^{-1}$. Then the joint density must be of the form

$$f(x,y) \propto \exp[-(Bx+Cy+Dxy)], \ x > 0, \ y > 0,$$

with parameters B,C, and D being positive. In this case, even though the joint density is a maximum entropy density subject to moment constraints on X, Y and XY, the marginal density of X is proportional to $(Dx+C)^{-1}\exp(-Bx)$ (see [1], page 19). This is not the exponential density that maximizes entropy with positive real line as support and the usual constraint on the mean E(X). Further, the correlation coefficient between X and Y is always non-positive.

In summary, one can say that there are two approaches of obtaining bivariate maximum entropy densities. One usual approach is to start with marginal maximum entropy densities and get joint maximum entropy densities by imposing constraints on bivariate moments. The other is to start with conditional maximum entropy densities and construct a joint density. Theorem 1 proves that the second

approach also leads to a maximum entropy density. The examples above bring out the differences in the two approaches.

As a final remark, it may be pointed out that the proof of Theorem 1 above, as it follows directly from the results of Arnold <u>et.al.</u> [1], is not at all intricate. On the other hand, an alternative proof, without the use of their results, seems quite difficult.

**References and Notes**

[1] Arnold, B.C.; Castillo, E.; Sarabia, J. *Conditionally Specified Distributions*. *Lecture Notes in Statistics*; No. 73, Springer-Verlag: New York, NY. 1992.

[2] Gokhale, D.V. "Maximum entropy characterizations of some distributions." In *Statistical Distributions in Scientific Work*; Patil, G.P.; Kotz, S.; Ord, J.K. Eds., D. Reidel Publishing: Boston, MA, 299-304, 1975.

[3] Kagan, A.M.; Linnik, Yu.V.; Rao, C.R. *Characterization Problems in Mathematical Statistics*; J. Wiley: New York, NY. 1973.