Distributions with Fixed Marginals and Related Topics
IMS Lecture Notes – Monograph Series Vol. 28, 1996

A Distribution with Given Marginals And Given Regression Curve

By C. M. Cuadras
Universitat de Barcelona

Given two cdfs F and G, a cdf H with a linear regression curve and belonging to the Fréchet class $\mathbb{F}(F, G)$ was obtained (Cuadras, 1992). This paper extends this construction to the nonlinear case. If φ is a monotone nonlinear function, satisfying some restrictions (e.g., Vitale, 1979), a distribution H_φ belonging to $\mathbb{F}(F, G)$ is constructed, where the regression curve is a linear expression in φ. The cases where φ is increasing or decreasing are studied separately. The general case is obtained by means of mixtures and convex sums. Some consequences are: approximation of a bivariate distribution by another linear regression; bounds for the Hoeffding correlations; and the possibility of using this construction to test nonlinear regression procedures and methods of estimation of the regression curve. Some inequalities concerning the extremal correlations are also obtained and a multivariate extension is proposed.

1. Introduction. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two second-order random vectors of dimensions m and n and cdf's F and G respectively. Cuadras (1992) obtained a family $\mathbb{F}(F, G; \boldsymbol{R}_{xy})$ of joint distributions with given marginals F and G and given intercorrelation matrix $\boldsymbol{R}_{xy}$. For this family the regression curve $\boldsymbol{Y}/\boldsymbol{X}$ and all bivariate regressions Y/X are linear.

In this paper we construct the family $\mathbb{F}(F, G; \varphi)$ of joint distributions H_φ having a regression curve

$$\boldsymbol{y} = m(\boldsymbol{x}) = E(\boldsymbol{Y}/\boldsymbol{X} = \boldsymbol{x}),$$

which equals a given function $\varphi(\boldsymbol{x})$ up to an affine transformation

$$m(\boldsymbol{x}) = \alpha\phi(\boldsymbol{x}) + \beta. \tag{1}$$

The function φ satisfies the conditions stated by Vitale (1979). This construction is different from the approach of Arnold, Castillo and Sarabia

AMS 1991 Subject Classification: Primary 62E10; Secondary 62J02

Key words and phrases: Fréchet classes, mixture of distributions, nonlinear regression, extremal correlations.

(1993). To specify $F(\boldsymbol{x},\boldsymbol{y})$ they use the conditional distribution $F_{X/Y}(\boldsymbol{x}/\boldsymbol{y})$ and a compatible regression function $E(\boldsymbol{Y}/\boldsymbol{X}=\boldsymbol{x})=m(\boldsymbol{x})$.

The regression of Y on X is linear for most copulas (Hutchinson and Lai, 1991). However for marginals other than uniform distributions, the regression curves do not generally have simple analytic forms. An important example is the Plackett family of distributions when X and Y are normal variables (Mardia, 1967). Furthermore, to our knowledge, there is very little research in the case of multivariate marginals. Finally, to generate data for testing statistical regression models is another motivation of this paper.

2. Bivariate Distributions with an Increasing Regression Curve. Here X,Y are continuous r.v.'s, φ is a function with positive derivative $\varphi'(x)>0$, $x\in\mathbb{S}_x$, and $\mathbb{S}_x\subseteq\psi(\mathbb{S}_y)$, where $\psi=\varphi^{-1}$ and $\mathbb{S}_x,\mathbb{S}_y$ are the corresponding supports. Means and variances will be denoted by μ_x,μ_y,σ_x^2 and σ_y^2. We introduce the family $\mathbb{F}(F,G;\varphi)$ of bivariate distributions H_θ^+ defined by the mixture

$$H_\theta^+(x,y)=\theta F(\min\{x,\psi(y)\})+(1-\theta)F(x)J_\theta(y),\quad 0\le\theta\le\theta^+\le 1, \tag{2}$$

where, for $\theta\neq 1$,

$$J_\theta(y)=[G(y)-\theta F(\psi(y))]/(1-\theta)\quad 0\le\theta\le\theta^+,$$

is a cdf and

$$\theta^+=\sup\{\theta|0\le\theta<1,\quad J_\theta(y)\quad\text{is a cdf}\}.$$

It is easily proved that $H(x,\infty)=F(x), H(\infty,y)=G(y)$, and that the conditional distribution of Y given $X=x$ is

$$F(y/x)=\theta\delta_{\{\psi(y)\ge x\}}+G(y)-\theta F(\psi(y)),$$

where δ is the indicator function.

Note that it can happen that J_θ is a cdf only for $\theta=0$ (e.g., $F(x)=G(x)=x$ and $\psi(y)=(2y-1)^{1/3}$). Then the variables are independent and the regression curve is a particular case of (1) for $\alpha=0$. Other values of θ, outside the interval $[0,1]$ but providing a cdf, are not considered here.

Taking the expectation

$$E_F[\varphi(X)]=\int ydF(\psi(y))=\int\varphi(x)dF(x),$$

we obtain the regression curve

$$m(x)=\mu_y+\theta(\varphi(x)-E_F[\varphi(X)]), \tag{3}$$

which is linear in $\varphi(x)$. Some properties of this family, which generalizes the convex combination of independence and Fréchet upper bound, are presented in the sequel.

1) The correlation ratio, defined as $\text{var}(E\{Y \mid X\})/\text{var}(Y)$, is given by

$$\eta^2 = \theta^2 \text{ Var }_F[\varphi(X)]/\sigma_y^2.$$

Thus $\eta^2 = 0$ iff $\theta^2 = 0$ (implying that X and Y are stochastically independent) and $\eta^2 = 1$ iff $\theta^2 = 1$ (implying that $Y = \varphi(X)$, i.e., the relation $F(X) = G(Y)$, holds a.s.).

2) The correlation coefficient is given by

$$\rho(X,Y) = \theta \text{ Cov } (X, \varphi(X))/(\sigma_x \sigma_y). \tag{4}$$

3) For $\psi = F^{-1} \circ G$ we find $\theta^+ = 1$ and (1) yields

$$H_\theta^+(x,y) = \theta F(\min\{x, F^{-1} \circ G(y)\}) + (1-\theta)F(x)G(y), \quad 0 \leq \theta \leq 1,$$

which contains the upper Fréchet bound ($\theta = 1$) and the independence case ($\theta = 0$).

4) The cdf (2) has a singular part with mass concentrated on the curve $(x, \varphi(x)), x \in S_x$. In fact, $P[Y = \varphi(X)] = \theta$.

5) Suppose that X and Y have densities f and g with respect to Lebesgue measure. Then, by differentiating $J_\theta(y)$,

$$g(y) - \theta f(\psi(y))\psi'(y) > 0, \quad y \in \mathbb{S}_y,$$

hence

$$\theta^+ = \text{ess} \inf_{y \in \mathbb{S}_y} \{g(y)/f(\psi(y))\psi'(y)\}.$$

Note that both $g(y)$ and $f(\psi(y))\psi'(y)$ are densities. Thus $0 \leq \theta^+ \leq 1$.

3. Bivariate Distributions with a Decreasing Regression Curve. Suppose that $\varphi'(x) < 0, x \in \mathbb{S}_x$. The family $\mathbb{F}(F, G; \varphi)$ consists of the bivariate distributions H_θ^- defined by the mixture

$$H_\theta^-(x,y) = \theta[F(x) - F(\psi(y))] + (1-\theta)F(x)K_\theta(y), \quad 0 \leq \theta \leq \theta^- \leq 1, \tag{5}$$

where, for $\theta \neq 1$,

$$K_\theta(y) = [G(y) - \theta[1 - F(\psi(y))]]/(1-\theta), \quad 0 \leq \theta \leq \theta^-,$$

is a cdf and

$$\theta^- = \sup\{\theta | 0 \leq \theta < 1, \quad K_\theta(y) \quad \text{is a cdf}\}.$$

It can be shown that the marginals of H_θ^- are F and G, and the regression curve is still expressed by (3), a linear expression in the decreasing curve $\varphi(x)$. This family generalizes the convex combination of independence and the Fréchet lower bound. Moreover, note that K_θ may be a cdf only for $\theta = 0$.

In the absolutely continuous case the value for θ^- is given by

$$\theta^- = \operatorname{ess} \inf_{y \in \mathbb{S}_y} \{-g(y)/f(\psi(y))\psi'(y)\}.$$

4. Approximating a Bivariate Distribution by a Linear One. Assume X and Y are standardized to mean 0 and variance 1. Suppose that we want to approximate the joint cdf H by a cdf L_θ having the same marginals and for which the regression of Y on X is linear. Choosing $\varphi(x) = \alpha x$, $\alpha > 0$, we obtain the bivariate cdf

$$L_\theta(x, y) = \theta F(\min\{x, y/\alpha\}) + (1 - \theta)F(x)J_\theta(y), \quad 0 \leq \theta \leq \theta_\alpha^+,$$

where $J_\theta(y)$ is a cdf and, when absolutely continuous,

$$\theta_\alpha^+ = \operatorname{ess} \inf_{y \in \mathbb{S}_y} \{\alpha g(y)/f(y/\alpha)\}.$$

The correlation coefficient for L_θ is given by

$$\rho(\alpha, \theta) = E_{L_\theta}(XY) = \int xy dL_\theta(x, y) = \int \theta xy dF(\min\{x, y/\alpha\})$$
$$+ (1 - \theta) \int x dF(x) \int y dJ_\theta(y) = \alpha\theta.$$

An approximating criterion could be to choose α and θ such that $\alpha\theta$ is as close as possible to the correlation coefficient $\rho = E_H(XY)$.

The approximation of H by a cdf with linear decreasing regression can be similarly constructed. We next present an application.

The maximum Hoeffding correlation is given by

$$\rho^+ = \int xy dH^+(x, y) = \int_0^1 F^{-1}(u)G^{-1}(u)du,$$

where $F^{-1}(u) = \inf\{x, F(x) \geq u\}$ and $H^+(x, y) = \min\{F(x), G(y)\}$ is the upper Fréchet bound. The maximum value for $\rho(\alpha, \theta) = \alpha\theta$ is given by

$$\rho_+ = \sup_{\alpha > 0}\{\alpha \max\{\operatorname{ess} \inf_{y \in \mathbb{S}_y} \{\alpha g(y)/f(y/\alpha)\}, \operatorname{ess} \inf_{y \in \mathbb{S}_x} \{\alpha f(y)/g(y/\alpha)\}\}\}.$$

As $0 \leq \rho_+ \leq \rho^+ \leq 1$, ρ_+ can be used as an approximation to ρ^+. Note that $\rho_+ = \rho^+ = 1$ if $F = G$ and $\rho_+ < \rho^+ < 1$ if $F \neq G$.

The minimum Hoeffding correlation coefficient is given by

$$\rho^- = \int xydH^-(x,y) = \int_0^1 F^{-1}(u)G^{-1}(1-u)du,$$

where $H^-(x,y) = \max\{F(x)+G(y)-1,0\}$ is the lower Fréchet bound. By considering $\varphi(x) = \alpha x$, $\alpha < 0$, we obtain

$$\rho_- = \inf_{\alpha<0}\{\alpha \max\{\operatorname{ess\,inf}_{y\in\mathbb{S}_y}\{-\alpha g(y)/f(y/\alpha)\}, \operatorname{ess\,inf}_{y\in\mathbb{S}_x}\{-\alpha f(y)/g(y/\alpha)\}\}\},$$

which satisfies $-1 \le \rho^- \le \rho_- \le 0$, providing an approximation to ρ^-.

Further, for distributions with any mean and variance we have

$$\rho_L(\alpha,\theta) = \alpha\theta(\sigma_x/\sigma_y).$$

As a consequence, assuming $\sigma_x \ge \sigma_y > 0$, the following inequalities hold:

$$0 \le \operatorname*{ess\,inf}_{x\in\mathbb{S}_y}\{g(x)/f(x)\} \le (\sigma_y/\sigma_x)\rho^+ \le (\sigma_y/\sigma_x) \le 1;$$

$$0 \le \operatorname*{ess\,inf}_{x\in\mathbb{S}_y}\{g(x)/f(-x)\} \le -(\sigma_y/\sigma_x)\rho^- \le (\sigma_y/\sigma_x) \le 1.$$

More generally, we could use (4) and seek an appropriate φ to obtain an approximation to the extremal correlations, the computation of which is sometimes difficult. Specific calculations were obtained by Moran (1967, gamma marginals), DeVeaux (1976, Tech. Report, Stanford University, log-normal marginals), Fujita (1979, binary marginals). Recently, Shih and Huang (1992) combined different marginals and approximated some extremal correlations by simulation, and Cuadras and Fortiana (1994) made some numerical computations and presented a statistical application.

5. Bivariate Distributions with a General Regression Curve. Let H_0 and H_1 be cdfs belonging to the Fréchet class $\mathbb{F}(F,G)$, with regression curves $m_0(x)$ and $m_1(x)$ respectively. It is straightforward that the mixture

$$H_\lambda = (1-\lambda)H_0 + \lambda H_1, \quad 0 \le \lambda \le 1,$$

also belongs to $\mathbb{F}(F,G)$ and has the regression curve

$$m_\lambda(x) = (1-\lambda)m_0(x) + \lambda m_1(x), \quad 0 \le \lambda \le 1.$$

Suppose that $\varphi_0(x)$ and $\varphi_1(x)$ are two functions, increasing and decreasing respectively. Let

$$H(x,y) = (1-\lambda)H^+_{\theta_0}(x,y) + \lambda H^-_{\theta_1}(x,y), \quad 0 \le \lambda \le 1,$$

where $H^+_{\theta_0}$ is defined in (2), $H^-_{\theta_1}$ is defined in (5), and θ_0, θ_1 satisfy the condition that $J_{\theta_0}, K_{\theta_1}$ are cdf's. The regression curve for the mixture $H(x,y)$ is then

$$m(x) = \mu_y + (1-\lambda)\theta_0(\varphi_0(x) - E_F[\varphi_0(X)]) + \lambda\theta_1(\varphi_1(x) - E_F[\varphi_1(X)]),$$

which is linear in the convex combination

$$\varphi_\lambda(x) = (1-\lambda)\varphi_0(x) + \lambda\varphi_1(x), \quad 0 \le \lambda \le 1.$$

This construction is quite general. Let φ be any suitable function with domain $[a,b]$. If $|\varphi'(x)| \le A$ for $x \in (a,b)$, then $\varphi(x) = c(x) + d(x)$, where $c(x)$ is increasing and $d(x)$ is decreasing. Hence we only have to express, in an appropriate way, $\varphi(x) = (1-\lambda)\varphi_0(x) + \lambda\varphi_1(x)$, $0 \le \lambda \le 1$, and, by following the above construction, we can obtain a cdf H_φ belonging to $\mathbb{F}(F, G; \varphi)$.

6. Multivariate Distributions with a Given Regression Curve. Let $\boldsymbol{X} \sim F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^m$, $\boldsymbol{Y} \sim G(\boldsymbol{y})$, $\boldsymbol{y} \in \mathbb{R}^n$, be random vectors and $\varphi: \mathbb{R}^m \to \mathbb{R}^n$ a function such that $\varphi(\mathbb{S}_x) \subseteq \mathbb{S}_y$. We indicate by $F_\varphi(\boldsymbol{z})$ the cdf of $\boldsymbol{Z} = \varphi(\boldsymbol{X})$ and $H_\varphi(\boldsymbol{x}, \boldsymbol{y})$ the cdf of the random vector $(\boldsymbol{X}, \varphi(\boldsymbol{X}))$. Note that $H_\varphi \in \mathbb{F}(F, F_\varphi)$.

Let us introduce the mixture

$$H_\theta(\boldsymbol{x}, \boldsymbol{y}) = \theta H_\varphi(\boldsymbol{x}, \boldsymbol{y}) + (1-\theta)F(\boldsymbol{x})J_\theta(\boldsymbol{y}), \quad 0 \le \theta \le \theta^+,$$

where, for $\theta \ne 1$,

$$J_\theta(\boldsymbol{y}) = [G(\boldsymbol{y}) - \theta F_\varphi(\boldsymbol{y})]/(1-\theta), \quad 0 \le \theta \le \theta^+,$$

is a cdf and

$$\theta^+ = \sup\{\theta | 0 \le \theta < 1,\ J_\theta(\boldsymbol{y}) \quad \text{is a cdf}\}.$$

It is easily proved that $H_\varphi \in \mathbb{F}(F, G; \varphi)$. When φ is an affine transformation, Cuadras(1992) found an explicit expression for H_φ.

The results obtained from this approach, made explicit in the bivariate case, provide a proof of the following

THEOREM. *Let $\boldsymbol{X}, \boldsymbol{Y}$ be random vectors and φ a function as described above. A sufficient condition for a linear combination $m = \alpha\varphi + \beta$ to be a regression curve $\boldsymbol{Y}/\boldsymbol{X}$ is that $J_\theta(\boldsymbol{y}) = [G(\boldsymbol{y}) - \theta F_\varphi(\boldsymbol{y})]/(1-\theta)$ yields a cdf for some $0 \le \theta < 1$, where F_φ is the cdf of $\varphi(\boldsymbol{X})$.*

The above construction is particularly interesting when F, G are absolutely continuous with densities f, g (Lebesgue measure), $m = n = p$, and

$\boldsymbol{y} = \varphi(\boldsymbol{x})$ is one-to-one satisfying the change of variables condition for multiple integrals. Then the density of H_θ with respect to a measure μ is given by

$$h_\theta(\boldsymbol{x}, \boldsymbol{y}) = \theta f(\boldsymbol{x})\delta_{\{y=\varphi(x)\}} + (1-\theta)f(\boldsymbol{x})j_\theta(\boldsymbol{y}), \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p,$$

where

$$j_\theta(\boldsymbol{y}) = [g(\boldsymbol{y}) - \theta f(\varphi^{-1}(\boldsymbol{y})|\boldsymbol{J}|]/(1-\theta), \quad 0 \leq \theta \leq \theta^+,$$

$\boldsymbol{J} = (\partial \boldsymbol{x}/\partial \boldsymbol{y})$ is the Jacobian of $\boldsymbol{x}$ with respect to $\boldsymbol{y}$, and

$$\theta^+ = \operatorname{ess} \inf_{y \in \mathbb{S}_y} \{g(\boldsymbol{y})/f(\varphi^{-1}(\boldsymbol{y}))|\boldsymbol{J}|\}.$$

The (Lebesgue) measure μ can be expressed as $\mu = \mu_\varphi + \mu^{2p}$, where μ_φ is the measure concentrated on the surface $\boldsymbol{y} = \varphi(\boldsymbol{x})$ and μ^{2p} is the measure on $\mathbb{R}^{2p}$.

7. Discussion. We have constructed a family of multivariate distributions given the marginals and a compatible regression curve. This distribution concentrates mass on this curve and hence has a singular part. This drawback limits the applications when dealing with real data.

We have obtained a sufficient condition for a function to be a regression curve and we have also derived some inequalities concerning the extremal correlations.

Finally, given a function φ and the marginal distributions, this construction may be used to generate samples for testing methods of estimation of a regression curve and alternative methods of nonlinear regression (Cuadras and Arenas, 1990; Cuadras and Fortiana, 1993, 1995). This is possibly the principal application of this paper.

References

ARNOLD, B. C., CASTILLO, E. and SARABIA, J. M. (1993). Conditionally specified models: Structure and Inference. In: *Multivariate Analysis: Future Directions 2* (C. M. Cuadras and C. R. Rao, eds.). Elsevier Science Pub., Amsterdam, 441–450.

CUADRAS, C. M. (1992). Probability distributions with given multivariate marginals and given dependence structure. *J. Mult, Anal.* **42 (1)**, 51–66.

CUADRAS, C. M. and ARENAS, C. (1990). A distance based model for prediction with mixed data. *Comm. Statist. Theory Methods,* **19 (6)**, 2261–2279.

CUADRAS, C. M. and FORTIANA, J. (1993). Continuous metric scaling and prediction. In: *Multivariate Analysis: Future Directions 2* (C. M. Cuadras and C. R. Rao, eds.). Elsevier Science Pub., Amsterdam, 47–66.

CUADRAS, C. M. and FORTIANA, J. (1994). Ascertaining the underlying distribution of a data set. In: *Selected Topics on Stochastic Modelling* (R. Gutierrez and M. J. Valderrama, eds.). World Scientific, Singapore, 223–230.

CUADRAS, C. M. and FORTIANA, J. (1995). A continuous metric scaling solution for a random variable. *J. Mult. Anal.* **52 (1)**, 1–14.

FUJITA, K. (1979). The range of correlation coefficients obtainable from $m \times n$ correlation tables with given marginal distributions. *Sci. Rep. Fac. Ed. Gifu Univ. Natur. Sci.* **7 (3)**, 394–406.

HUTCHINSON, T. P. and LAI, C. D. (1991). *The Engineering Statistician's Guide to Continuous Bivariate Distributions.* Rumsby Scient. Pub., Adelaide.

MARDIA, K. V. (1967). Some contributions to contingency–type bivariate distributions. *Biometrika* **54**, 235–239.

MORAN, P. A. P. (1967). Testing for correlation between non-negative variates. *Biometrika* **54 (3), (4)**, 385–394.

SHIH, W. J. and HUANG, W. M. (1992). Evaluating correlations with proper bounds. *Biometrics* **48 (4)**, 1207–1213.

VITALE, R. A. (1979). Regression with given marginals. *Ann. Statist.* **7 (3)**, 653–658.

C. M. CUADRAS
DEPART. D'ESTADISTICA
UNIVERSITAT DE BARCELONA
DIAGONAL 645, 08028 BARCELONA, SPAIN
carlesm@porthos.bio.ub.es