GRAPHICAL STUDY OF DETERMINING PLANE PROBABILITY DISTRIBUTIONS WITH GIVEN MARGINS
Translated Document

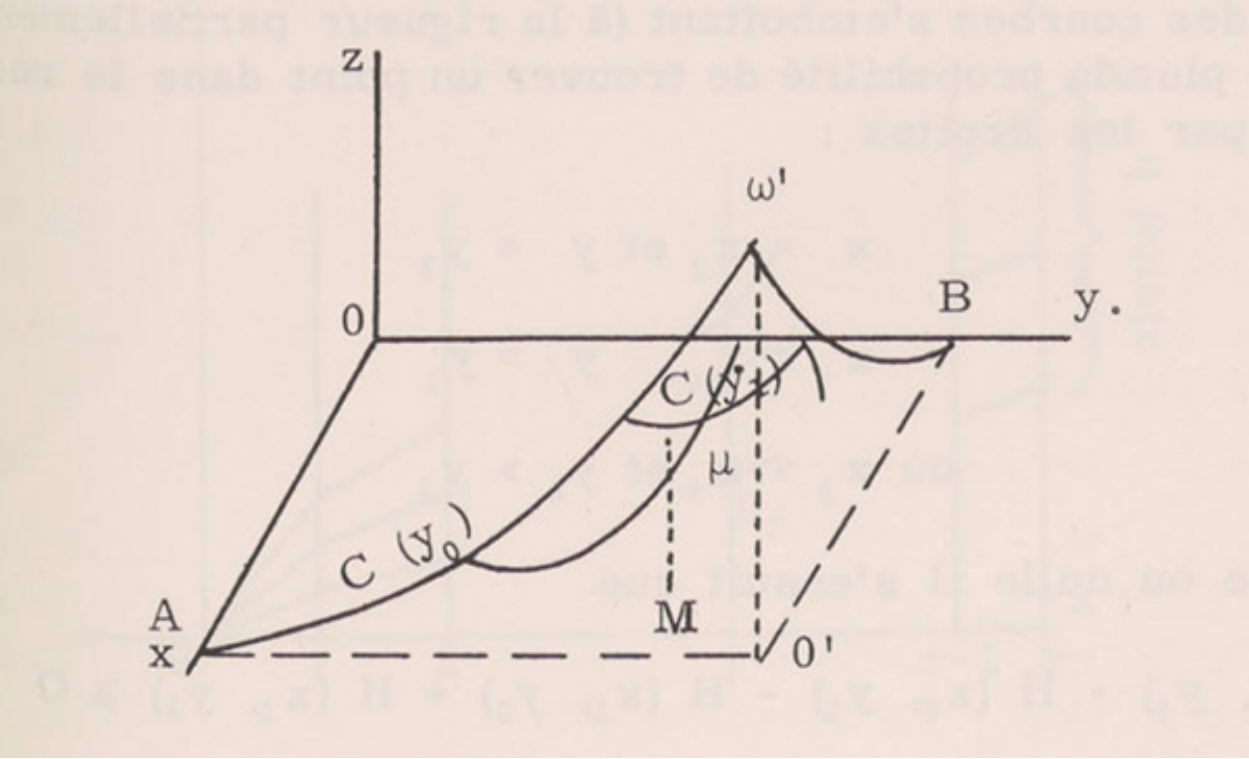
A. Nataf
Faculty of Sciences of Caen

In a previous note we indicated a very general method for determining probability distributions in R^n where the n marginal laws on each of the axes are assumed to be known

In this note we will limit ourselves to R^2 and indicate some geometric properties of such distributions $H(x, y)$ and then apply them to the more purely statistical problem of determining correlation tables whose margins are given.

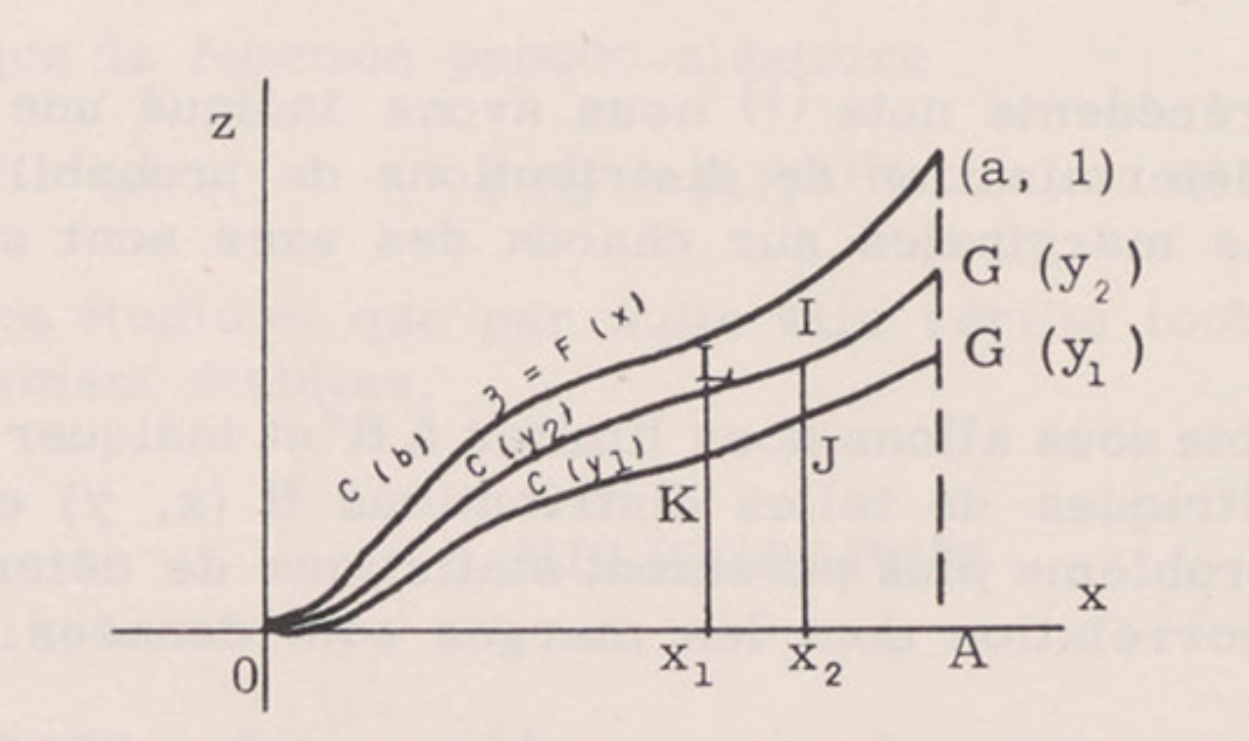
According to the previous note, we can reduce the problem by making two separate changes of variables on the x and y coordinates to the case where the entire distribution is contained within a rectangle of x, O_y with vertices, for example:

the origin and $A(a, o)$, $B(o, b)$ and $O'(a, b)$



To each point $M(x, y)$ of this rectangle, we associate the point $P(z, x, y)$ in the space where $z = H(x, y)$; P describes a surface, an indicator of $H(x, y)$. Thus, to O' corresponds ω' with a unit coordinate, and to O, O itself with a zero coordinate. We will determine S by determining the orthogonal projections onto $x\ O_z$, for example, of the sections $C(y_0)$ of S by the planes $y = c^{te} = y_0$ $(o \leq y_0 \leq b)$.

Knowledge of the marginal distributions $F(x)$ and $G(y)$ will translate, for $y = b$, into knowledge of the curve $C(b)$ with equation $z = F(x)$.



Furthermore, when projected onto the x-axis, the curves $C(y_0)$ will originate from a point with abscissa a and coordinate $G(y)$. Since, for the same value of x, $H(x, y)$ and $H(x, y_0)$ will coincide or be in the same order as y_1 and y_0, we see that the $C(y_0)$ curves will project along interlocking (or at most partially overlapping) curves. Moreover, the probability of finding a point in the rectangle IJKL defined by the lines:

$$x = x_2 \text{ and } y = y_2 \tag{1}$$

$$x = x_1 \text{ and } y = y_1 \tag{2}$$

$$\text{where } x_2 > x_1 \text{ and } y_2 > y_1 \tag{3}$$

being positive or zero, it follows that

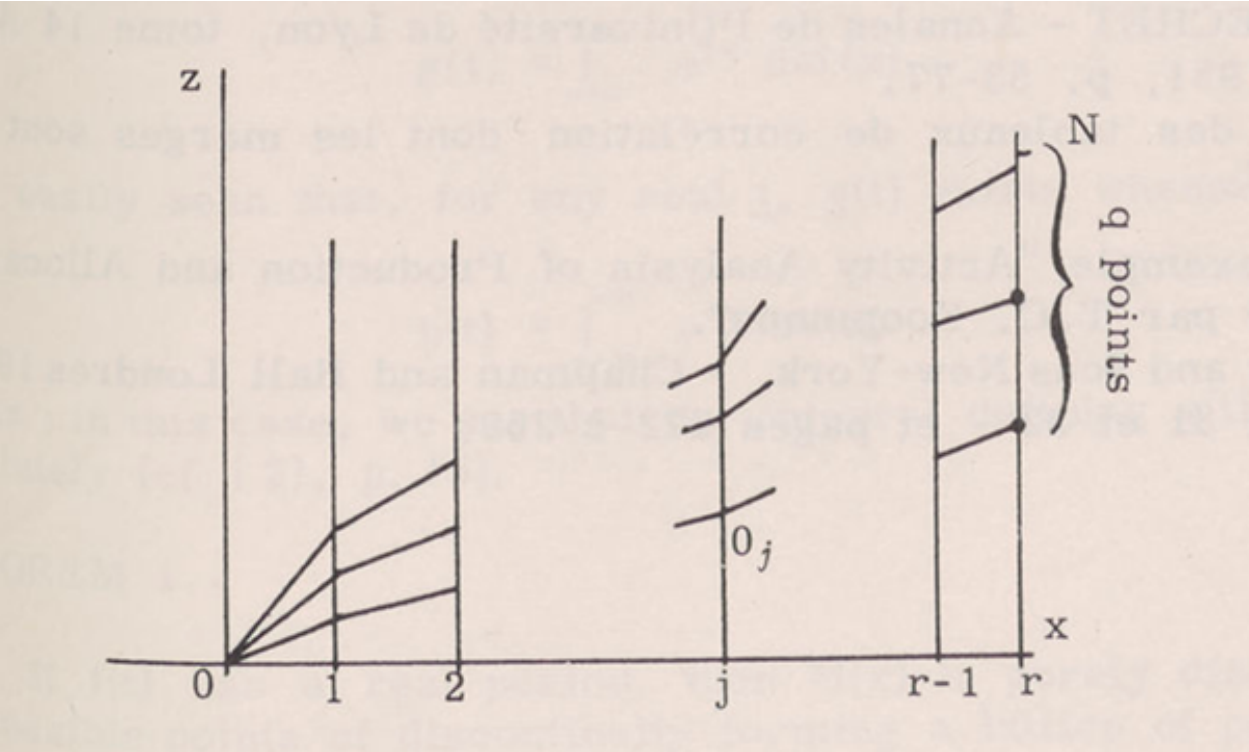
$$H(x_2, y_2) - H(x_2, y_1) - H(x_1, y_2) + H(x_1, y_1) \geq 0 \tag{4}$$

$$\text{which can be rewritten: } H(x_2, y_2) - H(x_2, y_1)H(x_2, y_2) - H(X_1, y_1) \tag{5}$$

and is interpreted as $IJ > LK$, or, given the non-decreasing nature of the functions represented by the CVs, by the fact that the slope of IL is greater than or at least equal to that of KJ.

Conversely, it is immediately apparent that if we can define a family of curves $\gamma(t)$ originating from O, having no other common point, converging on a point on the vertical axis of A and depending on a parameter t between 0 and 1, this parameter determining y by $G(y) = t$ or $y = G^{-1}(t)$, such that every chord of γ has a positive or zero slope, and that the slopes of the chords joining two pairs of points with the same abscissas on the curves $\gamma(t_1)$ and $\gamma(t_2)$ are parallel or of the same order of magnitude as t_1 and t_2, then the function $H(x, y) = z_t(x)$ (on the curve $t = G(y)$) defines a cumulative distribution function in two variables whose corresponding marginal distributions are $F(x)$ and $G(y)$.

This presentation is very intuitive and is likely useful in the study of families of distributions depending on parameters. It adapts very easily to the case where We consider two-dimensional statistical tables where each cell is filled with positive integers or zeros, allowing us to rediscover methods used by M. Frechet (1). If the table has q rows and r columns, we will consider, for example, in the x-axis, the r parallel lines Dj to the x-axis drawn through the r points on the x-axis with abscissas 1, 2, ..., r. We will scale the line x = r not from 0 to 1, but using points with integer coordinates, and we will draw the total distribution histogram representing $F(x)$, where F is no longer a function.



between 0 and 1 but an integer-valued function between 0 and N, where N is the total number of recorded events. The curves $C(y)$ will simply be replaced by polygonal contours joining points of integer elevation on the Dj. A set of contours (possibly reduced to a single point on Dr will always suffice if the slopes of the lines joining two pairs of points with the same abscissa on two contours are steeper the higher the contours belong to. It is easy to see that this method, applied to numerical cases, allows for a relatively rapid study, encompassing the problem at a glance. A systematic study, even in fairly complex cases, could be carried out by finding all possible contours immediately below the contour of $F(x)$, to which the other immediately below contours can be successively attached, and so on. The use of the bounds of $H(x, y)$ given by M. Frechet (1) greatly facilitates the study of these families of contours.

This algorithm can also, by simply removing the condition that the elevations be integers, allow the study of the so-called transportation problem in economics (2).

References

1 M. FRECHET- Annales de l'Université de Lyon, tome 14 A an née 1951, p. 53-77. "Sur des tableaux de corrélation dont les marges sont don nées".

2 Cf par exemple "Activity Analysis of Production and Allocation" (édité par T.C. Koopmans). Wiley and Sons New-York,- Chapman and Hall Londres 1951 pages 31 et 32- et pages 222 à 259