Canad. Math. Bull. Vol. **33** (1), 1990

DEPENDENT RANDOM VARIABLES WITH INDEPENDENT SUBSETS – II

BY

Y. H. WANG

ABSTRACT. In this paper, we consolidate into one two separate problems – dependent random variables with independent subsets and construction of a joint distribution with given marginals. Let $\mathbf{N} = \{1, 2, 3, \ldots\}$ and $\mathbf{X} = \{X_n;\ n \in \mathbf{N}\}$ be a sequence of random variables with non-degenerate one-dimensional marginal distributions $\{F_n;\ n \in \mathbf{N}\}$. An example is constructed to show that there exists a sequence of random variables $\mathbf{Y} = \{Y_n;\ n \in \mathbf{N}\}$ such that the components of a subset of $\mathbf{Y}$ are independent if and only if its size is $\leqq k$, where $k \geqq 2$ is a prefixed integer. Furthermore, the one-dimensional marginal distributions of $\mathbf{Y}$ are those of $\mathbf{X}$.

Introduction. The first counter example to show that three random variables may be pairwise independent, but not mutually independent, was discovered by S. N. Bernstein. His example can be briefly described as follows: Let X_1 and X_2 be two independent Bernoulli random variables with $P(X_i = 1) = 1/2 = 1 - P(X_i = 0)$ for $i = 1, 2$. Define $X_3 = 1 + X_1 + X_2 \pmod 2$. Then X_1, X_2 and X_3 are three identically distributed dependent, but pairwise independent, random variables.

In [5], Joffe constructed an infinite sequence of pairwise independent random variables of which any subsets of size greater than or equal to 3 are dependent. Later in [6], he demonstrated the existence of a set of n dependent random variables, each uniformly distributed over the set $\{0, 1, \ldots, p-1\}$, such that all subsets of size k or less are independent, where k, n and p *must* satisfy $2 \leqq k < n \leqq p + 1$ with p a prime number. If $k = p = 2$ and $n = 3$, Joffe's example is identical to Bernstein's.

In [9], we explicitly constructed a set of continuous as well as discrete dependent random variables of size $n \geqq 3$ such that all proper subsets are independent. If $n = 3$ and the random variables are of discrete type, our example reduces to that of Bernstein. It has been included in the book [8]. Another example along the line is in [2].

It is obvious that the joint distribution function of a set of random variables uniquely determines the marginal distributions of its components. Counter examples can be easily constructed to show that the converse is not true. Two excellent examples deserve special attention.

Received by the editors February 23, 1988 and, in revised form, December 12, 1988.

AMS 1980 Subject Classifications: 60E05.

Key words and phrases: Random variables, pairwise independence, independence, joint distribution, marginal distributions.

Let F and G be two given distribution functions. Hoeffding [4] showed that $H_0(x,y) = \max[F(x)+G(y)-1,0]$ and $H_1 = \min[F(x),G(y)]$ are joint distribution functions with F and G as marginals and every H with this property satisfies $H_0 \leqq H \leqq H_1$ and $\mathrm{Cov}(H_0) \leqq \mathrm{Cov}(H) \leqq \mathrm{Cov}(H_1)$. ($\mathrm{Cov}(H)$ denotes the covariance of the random variables with joint distribution H.) Hoeffding's results appeared a decade later in Fréchet [1]. Whitt [10] provided two simple proofs based on rearrangement theorems for Hoeffding's results. He also obtained interesting equivalent expressions for H_0 and H_1 as the joint distributions of two bivariate random vectors which are functions of *one* uniform random variable.

Morgenstern [7] constructed $H_\alpha(x) = F(x)G(y)\left[1+\alpha\left(1-F(x)\right)\left(1-G(y)\right)\right]$. Then $\{H_\alpha;\ -1 \leqq \alpha \leqq 1\}$ is an infinite family of joint distributions with F and G as marginals. If $\alpha = 0$, it reduces to the independent case. In general, α can be defined so that the joint distribution will have a specified correlation coefficient. Morgenstern's construction was later generalized by Gumbel [3] to any $n \geqq 2$ dimensional case. Let F_i be n distribution functions. Gumbel defined

$$H(x_1,\ldots,x_n) = \prod_{i=1}^{n} F_i(x_i)\left[1+\alpha\prod_{i=1}^{n}\left(1-F_i(x_i)\right)\right], \qquad -1 \leqq \alpha \leqq 1. \tag{1}$$

It is evident that the one-dimensional marginal distributions of H are F_i. If F_i are absolutely continuous with density functions f_i, then the corresponding density function h of H is easily seen to be

$$h(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_i(x_i)\left[1-\alpha\prod_{i=1}^{n}\left(2F_i(x_i)-1\right)\right]. \tag{2}$$

Strange as it may sound, there has been no attempts to consolidate the problem of finding a set of dependent random variables with independent subsets and the problem of constructing a joint distribution function with given marginals. If we let $n = 3$, then (1) becomes

$$\begin{aligned} H(x_1,x_2,x_3) &= F_1(x_1)F_2(x_2)F_3(x_3) \\ &\quad \times \left[1+\alpha\left(1-F_1(x_1)\right)\left(1-F_2(x_2)\right)\left(1-F_3(x_3)\right)\right]. \end{aligned} \tag{3}$$

For $\alpha \neq 0$, (3) provided an instantaneous example of three pairwise independent, but not mutually independent, random variables. It is a very simple example and places no restrictions on the three distribution functions F_1, F_2 and F_3.

In this paper, we shall generalize equation (1) and construct an infinite sequence of random variables such that the components of any proper subsets of it are independent if and only if their size is less than or equal to k, a prefixed positive integer $\geqq 2$, and they have prescribed one-dimensional marginal distributions.

The main result. Let $\mathbf{N} = \{1,2,3,\ldots\}$ be the set of positive integers and $\{X_n : n \in \mathbf{N}\}$ be a sequence of random variables with the non-degenerate one-dimensional marginal distributions $\{F_n : n \in \mathbf{N}\}$. Denote $\mathcal{F} = \{\mathbf{I}; \mathbf{I} \subset \mathbf{N}, |\mathbf{I}| < \infty\}$ to be the family of all finite subsets of $\mathbf{N}$. ($|\mathbf{I}|$ denotes the cardinality of $\mathbf{I}$). Define

$$\mathbf{A} = \left\{\alpha_i : i = 0, 1, 2, \ldots, \alpha_i \in (-1, 1)\backslash\{0\} \quad \text{and} \quad \sum_{i=1}^{\infty} |\alpha_i| \leqq 1\right\}.$$

THEOREM. *For a given sequence of random variables* $\mathbf{X} = \{X_n : n \in \mathbf{N}\}$, *there exists a sequence of random variables* $\mathbf{Y} = \{Y_n : n \in \mathbf{N}\}$ *such that the components of a subset of* $\mathbf{Y}$ *are independent if and only if its size is* $\leqq k$, *where* $k \geqq 2$ *is a fixed integer, and the one-dimensional marginal distribution functions of* $\mathbf{Y}$ *are those of* $\mathbf{X}$.

PROOF. Denote $G_i = 1 - F_i$ for all $i \in \mathbf{N}$. We construct a family of joint distribution functions $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$ as follows: For $\mathbf{I} = (i_1, \ldots, i_K) \in \mathcal{F}$, if $|\mathbf{I}| = K \leqq k$, we define

$$H_{\mathbf{I}}(x_{i_1}, \ldots, x_{i_K}) = \prod_{j=1}^{K} F_{i_j}(x_{i_j}), \tag{4a}$$

if $K = k + 1$,

$$H_{\mathbf{I}}(x_{i_1}, \ldots, x_{i_K}) = \prod_{j=1}^{K} F_{i_j}(x_{i_j}) \left[1 + \alpha_0 \prod_{j=1}^{K} G_{i_j}(x_{i_j})\right], \tag{4b}$$

if $K \geqq k + 2$,

$$H_{\mathbf{I}}(x_{i_1}, \ldots, x_{i_K}) = \prod_{j=1}^{K} F_{i_j}(x_{i_j}) \left\{1 + \Sigma\left[\prod_{s \in \mathbf{B}^c} \alpha_s \prod_{t \in \mathbf{B}} G_t(x_t)\right]\right\}, \tag{4c}$$

where $\mathbf{B}$ are subsets of $\mathbf{I}$ of size $k + 1$ and $\mathbf{B}^c = \mathbf{I}\backslash\mathbf{B}$, and the summation Σ is taken over all $\binom{K}{k+1}$ possible $\mathbf{B}$. In (4b) and (4c) α are elements of $\mathbf{A}$. To avoid cumbersome notations, the dependence of the family $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$ on the set $\mathbf{A}$ is suppressed.

It is easy to verify that the family of distribution functions $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$ defined by equations (4a)–(4c) are bonafide K-dimensional distribution functions and satisfy the following additional conditions:

(i) $H_{\mathbf{I}}(x_{i_1}, \ldots, x_{i_K}) = H_{\mathbf{I}'}(x_{i_{(1)}}, \ldots, x_{i_{(K)}})$, where $\mathbf{I}' = (i_{(1)}, \ldots, i_{(K)})$ is a permutation of $\mathbf{I}$.

(ii) $H_{\mathbf{J}}(x_{i_1}, \ldots, x_{i_m}) = H_{\mathbf{I}}(x_{i_1}, \ldots, x_{i_m}, \infty, \ldots, \infty)$, for all $\mathbf{J} = (i_1, \ldots, i_m)$ consisting of the first m ($1 \leqq m < K$) elements of $\mathbf{I}$, is again a member of the family $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$. (Note that its dependence on $\mathbf{A}$ is suppressed.)

(iii) $H_{\{i\}}(x) = F_i(x)$, for all $i = 1, 2, 3, \ldots$.

We now let $\mathbf{Y} = \{Y_n : n \in \mathbf{N}\}$ be a sequence of random variables governed by the family of distribution functions $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$, in such a way that for $\mathbf{I} = (i_1, \ldots, i_K) \in \mathcal{F}$, $H_{\mathbf{I}}$ is the joint distributions of $(Y_{i_1}, \ldots, Y_{i_K})$.

https://doi.org/10.4153/CMB-1990-004-6 Published online by Cambridge University Press

This completes our proof.

REMARK 1. If we require the family of distribution functions $\{F_i : i \in \mathbf{N}\}$ to be absolutely continuous with corresponding density functions $\{f_i : i \in \mathbf{N}\}$, then the density functions for the family $\{H_{\mathbf{I}} : \mathbf{I} \in \mathcal{F}\}$ are

$$h_{\mathbf{I}}(x_{i_1}, \dots, x_{i_K}) = \prod_{j=1}^{K} f_j(x_{i_j}), \tag{5a}$$

for $1 \leqq K \leqq k$,

$$h_{\mathbf{I}}(x_{i_1}, \dots, x_{i_K}) = \prod_{j=1}^{K} f_j(x_{i_j}) \left[1 - \alpha_0 \prod_{j=1}^{K} \left(2F_j(x_{i_j}) - 1\right) \right] \tag{5b}$$

for $K = k + 1$ and

$$h_{\mathbf{I}}(x_{i_1}, \dots, x_{i_K}) = \prod_{j=1}^{K} f_j(x_{i_j}) \left\{ 1 - \Sigma \left[\prod_{s \in \mathbf{B}^c} \alpha_s \prod_{t \in \mathbf{B}} \left(2F_t(x_t) - 1\right) \right] \right\}, \tag{5c}$$

for $K \geqq k + 2$, where $\mathbf{B}$, $\mathbf{B}^c$, Σ and α are as defined in the last section.

REMARK 2. Our construction (4*a*)–(4*c*) is not unique. To see this let $k = 2$ and $n = 4$, then both

$$\begin{aligned} F(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} F_i(x_i) \Bigg[1 + \sum_{i=1}^{4} \Bigg(\prod_{j \neq i} \alpha_i \left(1 - F_j(x_j)\right) \Bigg) \\ + \alpha \prod_{j=1}^{4} \left(1 - F_j(x_j)\right) \Bigg], \end{aligned} \tag{6a}$$

and

$$F(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} F_i(x_i) \left[1 + \sum_{i=1}^{4} \left(\prod_{j \neq i} \alpha_i \left(1 - F_j(x_j)\right) \right) \right], \tag{6b}$$

and any of their convex combinations will give the same marginal distributions of dimension 3 or less.

REFERENCES

1. M. Fréchet, *Sur les tableaux de corrélation dont les marges sont données*, Ann. Univ. Lyon, **14** A (1951) 53–77.

2. S. Geisser and N. Mantel, *Pairwise independence of jointly dependent variables*, Ann. Math. Statist. **33** (1962) 290.

3. E. J. Gumbel, *Distributions à plusieurs variables dont les marges sont données*, Comptes Rendus de l'Academie Sc., Paris, **246** (1958) 2717–2720.

4. W. Hoeffding, *Masstabinvariante Korrelationstheorie*, Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik des Universitat Berlin, **5** (1940) 199–233.

5. A. Joffe, *On a sequence of almost deterministic pairwise independent random variables*, Proc. Amer. Math. Soc. **29** (1971) 381–382.

6. ———, *On a set of almost deterministic k-independent random variables*, Ann. Probab. **2** (1974) 161–162.

7. D. Morgenstern, *Eingache Beispiele zweidimensionaler Verteilungen*, Mitteilungsbl. Math. Statist. **8** (1956) 234–235.

8. J. M. Stoyanov, *Counter Examples in Probability*, John Wiley & Sons, England, 1987.

9. Y. H. Wang, *Dependent random variables with independent subsets*, Amer. Math. Monthly, **86** (1979) 290–292.

10. W. Whitt, *Bivariate distribution with given marginals*, Ann. Statist. **4** (1976) 1280–1289.

Concordia University
Montreal, Québec, H3G 1M8