

MULTIVARIATE DISTRIBUTIONS WITH FIXED MARGINALS AND CORRELATIONS

MARK HUBER,* *Claremont McKenna College*
NEVENA MARIĆ,** *University of Missouri–St. Louis*

Abstract

Consider the problem of drawing random variates (X_1, \dots, X_n) from a distribution where the marginal of each X_i is specified, as well as the correlation between every pair X_i and X_j . For given marginals, the Fréchet–Hoeffding bounds put a lower and upper bound on the correlation between X_i and X_j . Any achievable correlation between X_i and X_j is a convex combination of these bounds. We call the value $\lambda(X_i, X_j) \in [0, 1]$ of this convex combination the *convexity parameter* of (X_i, X_j) with $\lambda(X_i, X_j) = 1$ corresponding to the upper bound and maximal correlation. For given marginal distributions functions F_1, \dots, F_n of (X_1, \dots, X_n) , we show that $\lambda(X_i, X_j) = \lambda_{ij}$ if and only if there exist symmetric Bernoulli random variables (B_1, \dots, B_n) (that is $\{0, 1\}$ random variables with mean $\frac{1}{2}$) such that $\lambda(B_i, B_j) = \lambda_{ij}$. In addition, we characterize completely the set of convexity parameters for symmetric Bernoulli marginals in two, three, and four dimensions.

Keywords: Multivariate; copula; correlation; Monte Carlo algorithm

2010 Mathematics Subject Classification: Primary 65C05

Secondary 68W20

1. Introduction

Consider the problem of simulating a random vector (X_1, \dots, X_n) with second moments where for all i the cumulative distribution function (CDF) of X_i is F_i , and for all i and j the correlation between X_i and X_j should be $\rho_{ij} \in [-1, 1]$. The correlation here is the usual notion

$$\text{corr}(X, Y) = \frac{\mathbb{E}\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\}}{\text{SD}(X)\text{SD}(Y)} = \frac{\mathbb{E}\{XY\} - \mathbb{E}\{X\}\mathbb{E}\{Y\}}{\text{SD}(X)\text{SD}(Y)}$$

for standard deviations $\text{SD}(X)$ and $\text{SD}(Y)$ that are finite, where \mathbb{E} is the expectation.

Let Ω denote the set of matrices with entries in $[-1, 1]$, all the diagonal entries equal 1, and are nonnegative definite. Then it is well-known that any correlation matrix (ρ_{ij}) must lie in Ω .

This problem, in different guises, appears in numerous fields: physics [16], engineering [11], ecology [4], and finance [12], to name just a few. Due to its applicability in the generation of synthetic optimization problems, it has also received special attention from the simulation community; see [8] and [9].

A variety of approaches exist for this well-studied problem. When the marginals are normal and the distribution is continuous with respect to the Lebesgue measure, this is just the problem of generating a multivariate normal with specified correlation matrix. The method in order to accomplish this (see, for example, [6, p. 223]) for any matrix in Ω is well known.

Received 27 June 2013; revision received 14 April 2014.

* Postal address: Claremont McKenna College, 850 Columbia Avenue, Claremont, CA 91711, USA.

Email address: autotomic@gmail.com

** Postal address: University of Missouri–St. Louis, 1 University Boulevard, St. Louis, MO 63121, USA.

For marginals that are not normal, the question is very much harder. A common method is to employ families of copulas (see, for example, [15]), but there are very few techniques that apply to general marginals. Instead, different families of copulas are typically applied to different marginal distributions.

Devroye and Letac [3] showed that if the marginals are beta distributed with equal parameters at least $\frac{1}{2}$ then when the dimension is three it is possible to simulate such a vector where the correlation is any matrix in Ω . This set of beta distributions includes the important case of uniform $[0, 1]$ marginals, but the authors have not been able to extend their technique to higher dimensions.

Chagnanty and Joe [1] characterized the achievable correlation matrices when the marginals are Bernoulli. When the dimension is three their characterization is easily checkable, in higher dimensions they provide a number of inequalities that grow exponentially in the dimension.

For the case of general marginals, in statistics there is a tradition of using transformations of multivariate normal vectors, which dates back to Mardia [14] and Li and Hammond [13]. This approach relies heavily on developing usable numerical methods. In this paper we approach the same problem using exclusively probabilistic techniques.

We show that for many correlation matrices the problem of simulating from a multivariate distribution with fixed marginals and specified correlation can be reduced to showing the existence of a multivariate distribution whose marginals are Bernoulli with mean $\frac{1}{2}$, and for each pair of marginals there is a specified probability that the pair takes on the same value. For $n = 2, 3, 4$, we are able to give necessary and sufficient conditions on those agreement probabilities in order for such a distribution to exist.

The convexity graph. Any two random variables X and Y have correlation in $[-1, 1]$, but if the marginal distributions of X and Y are fixed, it is generally not possible to build a bivariate distribution for any correlation in $[-1, 1]$. For instance, for X and Y both exponentially distributed, the correlation must lie in $[1 - \pi^2/6, 1]$. The range of achievable correlations is always a closed interval.

For two dimensions the method to find the minimum and maximum correlation is well known. This comes from the inverse transform method, which works as follows. First, given a CDF F , define the pseudoinverse of the CDF as

$$F^{-1} = \inf\{x : F(x) \geq u\}. \tag{1}$$

When U is uniform over the interval $[0, 1]$ (write $U \sim \text{unif}([0, 1])$), $F^{-1}(U)$ is a random variable with CDF F (see, for example, [2, p. 28]). Since U and $1 - U$ have the same distribution, both can be used in the inverse transform method. The random variables U and $1 - U$ are *antithetic* random variables. Of course $\text{corr}(U, U) = 1$ and $\text{corr}(U, 1 - U) = -1$, so these represent an easy way to obtain minimum and maximum correlation when the marginals are uniform random variables.

The following theorem comes from [7] and [10].

Theorem 1. (Fréchet–Hoeffding bound.) *For X_1 with CDF F_1 and X_2 with CDF F_2 , and $U \sim \text{unif}([0, 1])$,*

$$\text{corr}(F_1^{-1}(U), F_2^{-1}(1 - U)) \leq \text{corr}(X_1, X_2) \leq \text{corr}(F_1^{-1}(U), F_2^{-1}(U)).$$

In other words, the maximum correlation between X_1 and X_2 is achieved when the same uniform is used in the inverse transform method to generate both. The minimum correlation between X_1 and X_2 is achieved when antithetic random variates are used in the inverse transform method.

Definition 1. Consider the random variables X and Y with finite second moments, and CDF F_X and F_Y , respectively. For U uniform on $[0, 1]$, let $\rho^- = \text{corr}(F_X^{-1}(U)F_Y^{-1}(1 - U))$ and $\rho^+ = \text{corr}(F_X^{-1}(U)F_Y^{-1}(U))$. Then (by the Fréchet–Hoeffding bound) there is a unique $\lambda \in [0, 1]$ such that

$$\text{corr}(X, Y) = \lambda\rho^+ + (1 - \lambda)\rho^-.$$

Call $\lambda = \lambda(X, Y)$ the *convexity parameter* of X and Y .

Definition 2. Consider (X_1, \dots, X_n) with finite second moments, where each X_i has CDF F_i , and the correlation between X_i and X_j is ρ_{ij} . Then the complete graph on $\{1, \dots, n\}$ where edge $\{i, j\}$ has weight $\lambda_{ij} = \lambda(X_i, X_j)$ is the *convexity graph* of the distribution.

Let \mathcal{B}_n be the set of probabilities on $\{0, 1\}^n$ such that if $(B_1, \dots, B_n) \sim \mu$ where $\mu \in \mathcal{B}_n$, then $\mathbb{P}\{B_i = 1\} = \frac{1}{2}$ for all i .

Theorem 2. Let $(B_1, \dots, B_n) \sim \mu \in \mathcal{B}_n$. Then $\lambda(B_i, B_j) = \mathbb{P}\{B_i = B_j\}$ for all $i < j$. For all distribution functions F_1, \dots, F_n with second moments there exists a distribution for (X_1, \dots, X_n) such that for all i we have $\mathbb{P}\{X_i \leq x\} = F_i(x)$ and for all $i < j$ we have $\lambda(X_i, X_j) = \lambda(B_i, B_j)$, where \mathbb{P} is the probability measure.

Proof. For (B_i, B_j) with symmetric Bernoulli marginals, the value of either $\mathbb{P}\{B_i = B_j\}$ or $\text{corr}(B_i, B_j)$ (which is one-to-one with $\lambda(B_i, B_j)$) completely determines the bivariate distribution. It is then straightforward to verify that $\mathbb{P}\{B_i = B_j\} = \lambda(B_i, B_j)$.

Next, consider U uniform on $[0, 1]$ independent of (B_1, \dots, B_n) . Then $X_i = F_i^{-1}(UB_i + (1 - U)(1 - B_i))$ has the correct marginals and again it is straightforward to show $\lambda(X_i, X_j) = \lambda(B_i, B_j)$.

From Theorem 2, we immediately see a way to simulate from a distribution (X_1, \dots, X_n) with given convexity parameters in linear time, provided it is possible to simulate from a multivariate symmetric Bernoulli with the same convexity parameters. The next result characterizes when such a multivariate Bernoulli exists in two, three, and four dimensions, and provides necessary conditions for higher dimensions.

Theorem 3. Suppose that (B_1, B_2, \dots, B_n) are $\{0, 1\}$ random variables with mean $\frac{1}{2}$ for all i . When $n = 2$, it is possible to simulate (B_1, B_2) for any $\lambda_{12} \in [0, 1]$. When $n = 3$, it is possible to simulate (B_1, B_2, B_3) if and only if

$$1 + 2 \min\{\lambda_{23}, \lambda_{12}, \lambda_{13}\} \geq \lambda_{23} + \lambda_{12} + \lambda_{13} \geq 1.$$

When $n = 4$, it is possible to simulate (B_1, B_2, B_3, B_4) if and only if

$$1 - \frac{1}{2}\ell \leq \frac{1}{2}(u - 1),$$

where

$$\begin{aligned} \ell &= \min(\lambda_{14} + \lambda_{24} + \lambda_{13} + \lambda_{23}, \lambda_{14} + \lambda_{34} + \lambda_{12} + \lambda_{23}, \lambda_{24} + \lambda_{34} + \lambda_{12} + \lambda_{13}), \\ u &= \min_{\{i,j,k\}} \lambda_{ij} + \lambda_{jk} + \lambda_{ik}. \end{aligned}$$

The rest of the paper is organized as follows. In section 2 we present the proof of Theorem 3. In Section 2.2, the set of multivariate asymmetric Bernoulli distributions is linked to that of the symmetric Bernoulli distributions. We conclude with a discussion in Section 3

2. Proof of Theorem 3

2.1. The $n = 2$ and $n = 3$ cases

Lemma 1. For any $\lambda_{12} \in [0, 1]$, there exists a unique joint distribution on $\{0, 1\}^2$ such that (B_1, B_2) with this distribution has $B_1, B_2 \sim \text{bern}(\frac{1}{2})$, and $\mathbb{P}\{B_1 = B_2\} = \lambda_{12}$, where ‘bern’ denotes a Bernoulli trial.

Proof. Let $p_{ij} = \mathbb{P}\{B_1 = i, B_2 = j\}$. Then the equations that are necessary and sufficient to meet the distribution and convexity conditions are

$$p_{10} + p_{11} = \frac{1}{2}, \quad p_{01} + p_{11} = \frac{1}{2}, \quad p_{11} + p_{00} = \lambda_{12}, \quad p_{00} + p_{01} + p_{10} + p_{11} = 1.$$

This system of linear equations has full rank, so there exists a unique solution. Given there is a unique solution, it is easy to verify that the solution is

$$p_{00} = \frac{1}{2}\lambda_{12}, \quad p_{01} = \frac{1}{2}[1 - \lambda_{12}], \quad p_{10} = \frac{1}{2}[1 - \lambda_{12}], \quad p_{11} = \frac{1}{2}\lambda_{12}.$$

This provides an alternate algorithm to that found in [5] for simulating from bivariate distributions with correlation between $\rho_{1,2}^-$ and $\rho_{1,2}^+$.

Lemma 2. A random vector (B_1, B_2, B_3) with $B_i \sim \text{bern}(\frac{1}{2})$ exists (and is possible to simulate from in a constant number of steps) if and only if the convexity graph satisfies

$$1 \leq \lambda_{23} + \lambda_{12} + \lambda_{13} \leq 1 + 2 \min\{\lambda_{12}, \lambda_{13}, \lambda_{23}\}$$

Proof. Let $p_{ijk} = \mathbb{P}\{B_1 = i, B_2 = j, B_3 = k\}$. The first condition is $\sum_{i,j,k} p_{i,j,k} = 1$. There are three conditions from the marginals:

$$\sum_{j,k \in \{0,1\}} p_{1jk} = \frac{1}{2}, \quad \sum_{i,k \in \{0,1\}} p_{i1k} = \frac{1}{2}, \quad \sum_{ij \in \{0,1\}} p_{ij1} = \frac{1}{2},$$

and three conditions from the correlations

$$\sum_{k \in \{0,1\}} p_{00k} + p_{11k} = \lambda_{12}, \quad \sum_{j \in \{0,1\}} p_{0j0} + p_{1j1} = \lambda_{13}, \quad \sum_{i \in \{0,1\}} p_{i00} + p_{i11} = \lambda_{23}.$$

To obtain eight equations, suppose that $p_{111} = \alpha$.

This eight-by-eight system of equations has full rank, so there is a unique solution. It is easy to verify that the solutions are

$$\begin{aligned} p_{000} &= \frac{1}{2}(\lambda_{12} + \lambda_{13} + \lambda_{23} - 1) - \alpha, & p_{100} &= \frac{1}{2}(1 - (\lambda_{12} + \lambda_{13})) + \alpha, \\ p_{001} &= \frac{1}{2}(1 - (\lambda_{13} + \lambda_{23})) + \alpha, & p_{101} &= \frac{1}{2}\lambda_{13} - \alpha, \\ p_{010} &= \frac{1}{2}(1 - (\lambda_{12} + \lambda_{23})) + \alpha, & p_{110} &= \frac{1}{2}\lambda_{12} - \alpha, \\ p_{011} &= \frac{1}{2}\lambda_{23} - \alpha, & p_{111} &= \alpha. \end{aligned}$$

In order for this solution to yield probabilities, every p_{***} must lie in $[0, 1]$. Since $p_{111} = \alpha$, $\alpha \geq 0$. From the p_{011} , p_{101} , and p_{110} equations it follows that

$$0 \leq \alpha \leq \frac{1}{2} \min\{\lambda_{12}, \lambda_{23}, \lambda_{13}\}. \tag{2}$$

The p_{000} equation requires that

$$\alpha \leq \frac{1}{2}(\lambda_{13} + \lambda_{12} + \lambda_{23} - 1). \tag{3}$$

With these two conditions, with p_{001} , p_{010} , and p_{100} , we obtain the constraint

$$\frac{1}{2}(\lambda_{13} + \lambda_{12} + \lambda_{23} - \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\} - 1) \leq \alpha. \tag{4}$$

Combining (4) and (2), we obtain

$$\frac{1}{2}(\lambda_{13} + \lambda_{12} + \lambda_{23} - \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\} - 1) \leq \frac{1}{2} \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\}. \tag{5}$$

As long as an $\alpha \geq 0$ exists satisfying (3) and (5) holds, there exists a solution.

From this result, we see that not all positive definite correlation matrices are attainable with $\text{bern}(\frac{1}{2})$ marginals. For example, if $\lambda_{12} = \lambda_{13} = \lambda_{23} = \frac{3}{10}$ then $\rho_{12} = \rho_{13} = \rho_{23} = -\frac{2}{5}$. With diagonal entries 1, the ρ values form a positive definite graph, but it is impossible to build a multivariate distribution with $\text{bern}(\frac{1}{2})$ marginals with these correlations.

2.2. The $n = 4$ case: asymmetric Bernoulli distributions

To show the $n = 4$ case, it will be useful to understand the problem of drawing a multivariate Bernoulli (X_1, \dots, X_n) where $X_i \sim \text{bern}(p_i)$ where i is not necessarily $\frac{1}{2}$.

Lemma 3. *An n -dimensional multivariate Bernoulli distribution where the marginal of component i is $\text{bern}(p_i)$ and convexity graph Λ exists if and only if an $n + 1$ dimensional multivariate Bernoulli distribution exists with $\text{bern}(\frac{1}{2})$ marginals and convexity graph*

$$\left[\begin{array}{ccc|c} & & & p_1 \\ & \Lambda & & \vdots \\ & & & p_n \\ \hline p_1 & \dots & p_n & 1 \end{array} \right].$$

Proof. Suppose that such an $(n + 1)$ -dimensional distribution exists with $\text{bern}(\frac{1}{2})$ marginals and specified convexity graph. Let (B_1, \dots, B_{n+1}) be a draw from this distribution. Then set $X_i = \mathbf{1}_{\{B_i=B_{n+1}\}}$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. From the convexity graph it follows that $\mathbb{P}\{X_i = 1\} = p_i$, and for $i \neq j$, $\mathbb{P}\{X_i = X_j\} = \mathbb{P}\{B_i = B_j\} = \lambda_{ij}$.

Conversely, suppose that such an n -dimensional distribution with $\text{bern}(p_i)$ marginals exists. Let $B_{n+1} \sim \text{bern}(\frac{1}{2})$ independent of the X_i , and set $B_i = B_{n+1}X_i + (1 - B_{n+1})(1 - X_i)$. Then $\mathbb{P}\{B_i = 1\} = \frac{1}{2}p_i + \frac{1}{2}(1 - p_i) = \frac{1}{2}$, and $\mathbb{P}\{B_i = B_{n+1}\} = p_i$, the correct convexity parameter. Finally, for $i \neq j$,

$$\mathbb{P}\{B_i = B_j\} = \mathbb{P}\{X_i = X_j\} = \lambda_{ij}.$$

Lemma 3 can be used to finish the $n = 4$ case.

Lemma 4. A random vector (B_1, B_2, B_3, B_4) with $B_i \sim \text{bern}(\frac{1}{2})$ exists (and is possible to simulate in a constant number of steps) if and only if for

$$\ell = \min(\lambda_{14} + \lambda_{24} + \lambda_{13} + \lambda_{23}, \lambda_{14} + \lambda_{34} + \lambda_{12} + \lambda_{23}, \lambda_{24} + \lambda_{34} + \lambda_{12} + \lambda_{13}),$$

$$u = \min_{\{i,j,k\}} \lambda_{ij} + \lambda_{jk} + \lambda_{ik}$$

it is true that

$$1 - \frac{1}{2}\ell \leq \frac{1}{2}(u - 1).$$

Proof. By using Lemma 3, the problem is reduced to finding a distribution for (X_1, X_2, X_3) where $X_i \sim \text{bern}(\lambda_{i4})$ and the upper three-by-three minor of Λ is the new convexity graph. Just as in Lemma 2, this provides eight equations of full rank with a single parameter α . Letting $q_{ijk} = \mathbb{P}\{X_1 = i, X_2 = j, X_3 = k\}$, the unique solutions are

$$\begin{aligned} q_{000} &= \frac{1}{2}(\lambda_{12} + \lambda_{13} + \lambda_{23}) - \frac{1}{2} - \alpha, \\ q_{001} &= -\frac{1}{2}(\lambda_{14} + \lambda_{24} + \lambda_{13} + \lambda_{23}) + 1 + \alpha, \\ q_{010} &= -\frac{1}{2}(\lambda_{14} + \lambda_{34} + \lambda_{12} + \lambda_{23}) + 1 + \alpha, \\ q_{011} &= \frac{1}{2}(\lambda_{24} + \lambda_{34} + \lambda_{23}) - \frac{1}{2} - \alpha, \\ q_{100} &= -\frac{1}{2}(\lambda_{24} + \lambda_{34} + \lambda_{12} + \lambda_{13}) + 1 + \alpha, \\ q_{101} &= \frac{1}{2}(\lambda_{14} + \lambda_{34} + \lambda_{13}) - \frac{1}{2} - \alpha, \\ q_{110} &= \frac{1}{2}(\lambda_{14} + \lambda_{24} + \lambda_{12}) - \frac{1}{2} - \alpha, \\ q_{111} &= \alpha. \end{aligned}$$

All of these right-hand sides lie in $[0, 1]$ if and only if $1 - \frac{1}{2}\ell \leq \frac{1}{2}(u - 1)$, and α is chosen to lie in $[1 - \frac{1}{2}\ell, \frac{1}{2}(u - 1)]$.

As with the three-dimensional case, this proof can be used to simulate a four-dimensional multivariate symmetric Bernoulli: generate (X_1, X_2, X_3) using any $\alpha \in [1 - \frac{1}{2}\ell, \frac{1}{2}(u - 1)]$ and the q -distribution, then generate $B_4 \sim \text{bern}(\frac{1}{2})$, and then set B_i to be $B_4X_i + (1 - B_4)(1 - X_i)$ for $i \in \{1, 2, 3, 4\}$.

3. Conclusions

The Fréchet–Hoeffding bounds provide a lower and upper bound on the pairwise correlation between two random variables with given marginals. Hence, for higher dimensions the correlation matrix provides edge weights for a convexity graph whose parameters indicates where on the line from the lower to the upper bound the correlation lies. If it is possible to build a multivariate distribution with these convexities for marginals that are symmetric Bernoulli then it is possible to build a multivariate distribution with these convexities for arbitrary marginals. For two, three, and four dimensions, the set of convexity matrices that yield a symmetric Bernoulli distribution is characterized completely. For five or higher dimensions, every subset of three and four have these characterizations as necessary conditions.

References

- [1] Chaganty, N. R. and Joe, H. (2006). Range of correlation matrices for dependent Bernoulli random variables. *Biometrika* **93**, 197–206.
- [2] Devroye, L. (1986). *Nonuniform Random Variate Generation*. Springer, New York. 1986.

- [3] Devroye, L. and Letac, G. (2010). Copulas in three dimensions with prescribed correlations. Preprint. Available at <http://arxiv.org/abs/1004.3146v1>.
- [4] Dias, C. T. S., Samaranyaka, A. and Manly, B. (2008). On the use of correlated beta random variables with animal population modelling. *Ecological Model.* **215**, 293–300.
- [5] Dukic, V. M. and Marić, N. (2013). Minimum correlation in construction of multivariate distributions. *Phys. Rev. E* **87**, 032114.
- [6] Fishman, G. S. (1996). *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.
- [7] Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sect. A (3)* **14**, 53–77.
- [8] Henderson, S. G., Chiera, B. A. and Cooke, R. M. (2000). Generating ‘dependent’ quasi-random numbers. In *Proceedings of the Winter Simulation Conference, 2000*, Vol. 1 IEEE, New York, pp. 527–536.
- [9] Hill, R. R. and Reilly, C. H. (1994). Composition for multivariate random variables. In *Proceedings of the Winter Simulation Conference, 1994*, IEEE, New York, pp. 332–339.
- [10] (1940). Hoeffding, W. Masstabinvariante korrelatiostheorie. *Schriften Math. Inst. Univ. Berlin* **5**, 179–233.
- [11] Lampard, D. G. (1968). A stochastic process whose successive intervals between events form a first order Markov chain. I. *J. Appl. Prob.* **5**, pp. 648–668.
- [12] Lawrance, A. J. and Lewis, P. A. W. (1981). A new autoregressive time series model in exponential variables (NEAR (1)). *Adv. Appl. Prob.* **13**, 826–845.
- [13] Li, S. T. and Hammond, J. L. (1975). Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Trans. Systems Man Cybernetics* **5**, 557–561.
- [14] Mardia, K. V. (1970). A translation family of bivariate distributions and Fréchet’s bounds. *Sankhyā* **32**, 119–122.
- [15] Nelson, R. B. (2006). *An Introduction to Copulas*. Springer, New York.
- [16] Smith, O. E. and Adelfang, S. I. (1981) Gust model based on the bivariate gamma probability distribution. *J. Spacecraft Rockets* **18** 545–549.