Simulation of multivariate distributions with fixed marginals and correlations

Mark Huber and Nevena Marić

June 24, 2013

Abstract

Consider the problem of drawing random variates \((X_1, \ldots, X_n)\) 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 \(X_j\). Hence any achievable correlation can be uniquely represented by a convexity parameter \(\lambda_{ij} \in [0, 1]\) where 1 gives the maximum correlation and 0 the minimum correlation. We show that for a given convexity parameter matrix, the worst case is when the marginal distribution are all Bernoulli random variables with parameter 1/2 (fair 0-1 coins). It is worst case in the sense that given a convexity parameter matrix that is obtainable when the marginals are all fair 0-1 coins, it is possible to simulate from any marginals with the same convexity parameter matrix. In addition, we characterize completely the set of convexity parameter matrices for symmetric Bernoulli marginals in two, three and four dimensions.

1 Introduction

Consider the problem of simulating a random vector \((X_1, \ldots, X_n)\) 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 \(\rho_{ij} \in [0, 1]\). The correlation here is the usual notion

\[
\text{Cor}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\text{SD}(X)\text{SD}(Y)} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\text{SD}(X)\text{SD}(Y)},
\]

for standard deviations \(\text{SD}(X)\) and \(\text{SD}(Y)\) that are finite.

Let \(\Omega\) 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 \((\rho_{ij})\) must lie in \(\Omega\).
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 by the simulation community [9], [8].

It is a very well studied problem with a variety of approaches. When the marginals are normal and the distribution is continuous with respect to Lebesgue measure, this is just the problem of generating a multivariate normal with specified correlation matrix. It is well known how to accomplish this (see, for instance [6], p. 223) for any matrix in $\Omega$.

For marginals that are not normal, the question is very much harder. A common method is to employ families of copulas (see for instance [15]), but there are very few techniques that apply to general marginals. Instead, different families of copulas typically focus on 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 $\Omega$. This set of beta distributions includes the important case of uniform [0, 1] marginals, but they 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 3 their characterization is easily checkable, in higher dimensions they give a number of inequalities that grows exponentially in the dimension.

For the case of general marginals, in statistics there is a tradition of using transformations of multivariate normal vectors dating 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 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 matrix approach. 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 it is well known how to find the minimum and maximum correlation. These come 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 instance p. 28 of [2]). 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 $\text{Cor}(U, U) = 1$ and $\text{Cor}(U, 1-U) = -1$, so these represent an easy way to get minimum and maximum correlation when the marginals are uniform random variables.

The following theorem comes from work of Fréchet [7] and Hoeffding [10].

**Theorem 1** (Fréchet-Hoeffding bound). For $X_1$ with cdf $F_1$ and $X_2$ with cdf $F_2$, and $U \sim \text{Unif}(0, 1)$:

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

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.

Now consider $(X_1, \ldots, X_n)$, where each $X_i$ has cdf $F_i$, and the correlation between $X_i$ and $X_j$ is $\rho_{ij}$. Then let

$$\rho_{ij}^- = \text{Cor}(F_i^{-1}(U)F_j^{-1}(1-U)) \quad \text{and} \quad \rho_{ij}^+ = \text{Cor}(F_i^{-1}(U)F_j^{-1}(U)).$$

For the $\rho_{ij}$ to be achievable, it must be true that $\rho_{ij} \in [\rho_{ij}^-, \rho_{ij}^+]$, so there exists $\lambda_{ij} \in [0, 1]$ so that

$$\rho_{ij} = \lambda_{ij}\rho_{ij}^+ + (1 - \lambda_{ij})\rho_{ij}^-.$$ 

Given a correlation matrix $(\rho_{ij})$ we will refer to $(\lambda_{ij})$ as the convexity parameter matrix. Our main result is as follows.

**Theorem 2.** Suppose that for a convexity parameter matrix $\Lambda$, it is possible to simulate $(B_1, \ldots, B_n)$ where for all $i$, $\mathbb{P}(B_i = 1) = \mathbb{P}(B_i = 0) = 1/2$ and

$$(\forall i,j)(\mathbb{P}(B_i = B_j) = \lambda_{ij}).$$
Then for any specified set of marginals $F_1,\ldots,F_n$ it is possible to simulate in linear time from a multivariate distribution on $(X_1,\ldots,X_n)$ such that

$$(\forall i)(X_i \sim F_i) \text{ and } (\forall i,j)(\text{Cor}(X_i,X_j) = \lambda_{ij}\rho_{ij}^+ + (1 - \lambda_{ij})\rho_{ij}^-).$$

The next result characterizes when such a multivariate Bernoulli exists in two, three, and four dimensions, and gives necessary conditions for higher dimensions.

**Theorem 3.** Suppose $(B_1,B_2,\ldots,B_n)$ are random variables with $P(B_i = 1) = P(B_i = 0) = 1/2$ 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

$$1 - (1/2)\ell \leq (1/2)(u - 1)$$

where

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

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

The rest of the paper is organized as follows. In the next section the notion of a concurrence matrix is introduced, and it is shown that when the marginals are all Bernoulli with mean $1/2$, the concurrence matrix and convexity matrix are the same. Furthermore, given a multivariate Bernoulli $1/2$ with given convexity matrix, a straightforward algorithm gives the multivariate distribution with arbitrary marginals that has the same convexity matrix, proving Theorem 2. In Section 3, the $n = 2$ and $n = 3$ cases are shown. In Section 4, a link is made between symmetric Bernoulli marginals and asymmetric, which in turn gives the proof of the $n = 4$ case.

## 2 The algorithm

To present our algorithm, we begin by defining a matrix that measures the chance that any two components are equal.

**Definition 1.** For a random vector $(X_1,\ldots,X_n)$ the concurrence matrix $A = (a_{ij})$ is defined as $a_{ij} = P(X_i = X_j)$. 

4
Let $\text{Bern}(p)$ denote the Bernoulli distribution with parameter $p$, so for $X \sim \text{Bern}(p)$, $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$.

**Proposition 1.** For a random vector $(B_1, \ldots, B_n)$ with $B_i \sim \text{Bern}(1/2)$ for all $i$, the concurrence matrix and convexity matrix are the same matrix.

**Proof.** Let $(B_1, \ldots, B_n)$ be a random vector with marginals $\text{Bern}(1/2)$, and $i \neq j$ be elements of $\{1, \ldots, n\}$.

Let $1(\text{expression})$ denote the indicator function that is 1 if the expression is true and 0 otherwise. The inverse transform method for generating $\text{Bern}(1/2)$ random variables is $B = 1(U > 1/2)$ where $U \sim \text{Unif}([0, 1])$. From the Fréchet-Hoeffding bound (Theorem 1), the correlation between $B_i$ and $B_j$ can range anywhere from -1 to 1. Hence

$$\text{Cor}(B_i, B_j) = \lambda_{ij}(1) + (1 - \lambda_{ij})(-1) \Rightarrow \lambda_{ij} = (1/2)(\text{Cor}(B_i, B_j) + 1).$$

Now denote $\mathbb{P}(B_i = a, B_j = b)$ by $p_{ab}$. Then

$$\mathbb{P}(B_i = B_j) = p_{11} + p_{00} = a_{ij}$$
$$\mathbb{P}(B_i = 1) = p_{10} + p_{11} = 1/2$$
$$\mathbb{P}(B_j = 1) = p_{01} + p_{11} = 1/2$$
$$\sum_{a,b} \mathbb{P}(B_i = a, B_j = b) = p_{00} + p_{01} + p_{10} + p_{11} = 1$$

Adding the top three equations and then using the fourth gives

$$3p_{11} + p_{10} + p_{01} + p_{00} = a_{ij} + 1 = 2p_{11} + 1.$$ 

Hence $p_{11} = a_{ij}/2$.

That means

$$\text{Cor}(B_i, B_j) = \frac{\mathbb{E}[B_iB_j] - \mathbb{E}[B_i]\mathbb{E}[B_j]}{\text{SD}(B_1)\text{SD}(B_2)} = \frac{(1/2)a_{ij} - (1/2)(1/2)}{(1/2)(1/2)} = 2a_{ij} - 1,$$

so $a_{ij} = (1/2)(\text{Cor}(B_i, B_j) + 1)$, and $a_{ij} = \lambda_{ij}$.

Given the ability to generate from a multivariate Bernoulli with given convexity/concurrence matrix, generating from a multivariate with general marginals with the same convexity matrix turns out to be easy. Suppose the $n$ by $n$ convexity matrix $\Lambda = (\lambda_{ij})$ and cdf’s $F_1, \ldots, F_n$ are given.
Generating multivariate distributions with given convexity matrix

1) Draw $U$ uniformly from $[0,1]$.
2) Draw $(B_1, \ldots, B_n)$ ($B_i \sim \text{Bern}(1/2)$) with concurrence matrix $\Lambda = (\lambda_{ij})$.
3) For all $i \in \{1, \ldots, n\}$, let $X_i \leftarrow F_i^{-1}(UB_i + (1-U)(1-B_i))$.

That is, if $B_i = 1$ then choose $X_i$ with the inverse transform method and $U$, otherwise use the inverse transform method and $1-U$. Theorem 2 can now be restated in terms of this algorithm.

**Theorem 4.** The above algorithm generates $(X_1, \ldots, X_n)$ with $X_i \sim F_i$ for all $i$ and convexity matrix $\Lambda$.

**Proof.** Each $X_i$ can be written as $B_i F_i^{-1}(U) + (1-B_i) F_i^{-1}(1-U)$. That is, it is a mixture of $F_i^{-1}(U)$ and $F_i^{-1}(1-U)$ with weights equal to $1/2$.

Since $F_i^{-1}(U)$ and $F_i^{-1}(1-U)$ have the same distribution with cdf $F_i$, the random variable $X_i$ also shares this distribution. Hence they all have the same mean (call it $\mu_i$) and standard deviation (call it $\sigma_i$).

That means the correlations satisfy

$$\text{Cor}(X_i, X_j)\sigma_i\sigma_j = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]
= \mathbb{E}[\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]|B_i, B_j]
= \sum_{a\in\{0,1\}} \sum_{b\in\{0,1\}} \mathbb{P}(B_i = a, B_j = b) \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)|B_i = a, B_j = b].$$

Note when $B_i = B_j = 0$, then $X_i = F_i^{-1}(U)$ and $X_j = F_j^{-1}(U)$ which are maximally correlated. So

$$\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)|B_i = 0, B_j = 0] = \rho_{ij}^+ \sigma_i \sigma_j.$$

Following this same logic further gives

$$\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)|B_i = 1, B_j = 0] = \rho_{ij}^- \sigma_i \sigma_j$$
$$\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)|B_i = 0, B_j = 1] = \rho_{ij}^- \sigma_i \sigma_j$$
$$\mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)|B_i = 1, B_j = 1] = \rho_{ij}^+ \sigma_i \sigma_j$$

Hence

$$\text{Cor}(X_i, X_j) = \rho_{ij}^+ \mathbb{P}(B_i = B_j) + \rho_{ij}^- \mathbb{P}(B_i \neq B_j) = \rho_{ij}^+ a_{ij} + \rho_{ij}^-(1 - a_{ij}).$$
Theorem 2 is an immediate consequence.

The complete algorithm for generating multivariate distributions where component $i$ has cdf $F_i$ and $\text{Cor}(X_i, X_j) = \rho_{ij}$ is as follows.

<table>
<thead>
<tr>
<th>Generating multivariate distributions with given correlation matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) For every ${i,j} \subseteq {1,\ldots,n}$ do (where $A \sim \text{Unif}([0,1])$)</td>
</tr>
<tr>
<td>2) $\rho_{ij}^+ \leftarrow \text{Cor}(F_i^{-1}(A), F_i^{-1}(1-A))$, $\rho_{ij}^- \leftarrow \text{Cor}(F_i^{-1}(A), F_i^{-1}(1-A))$.</td>
</tr>
<tr>
<td>3) $\lambda_{ij} \leftarrow (\rho_{ij} - \rho_{ij}^-)/(\rho_{ij}^+ - \rho_{ij}^-)$</td>
</tr>
<tr>
<td>4) Draw $U \leftarrow \text{Unif}([0,1])$</td>
</tr>
<tr>
<td>5) Draw $(B_1,\ldots,B_n)$ $(B_i \sim \text{Bern}(1/2))$ with concurrence matrix $\Lambda = (\lambda_{ij})$.</td>
</tr>
<tr>
<td>6) For all $i \in {1,\ldots,n}$, let $X_i \leftarrow F_i^{-1}(UB_i + (1-U)(1-B_i))$.</td>
</tr>
</tbody>
</table>

Line 2 just generates the maximum and minimum correlation bounds from Theorem 1. Line 3 is set up so that $\rho_{ij} = \lambda_{ij}\rho_{ij}^+ + (1 - \lambda_{ij})\rho_{ij}^-$. Note that when the marginals for $i$ and $j$ are the same distribution, $\rho_{ij}^+ = 1$. However, even for equal marginals, the minimum correlation is often strictly greater than $-1$. See [5] and the references therein for the minimum correlation for a variety of common distributions.

This algorithm is only applicable when it is actually possible to simulate from the multivariate symmetric Bernoulli distribution with specified concurrence matrix. The next section gives necessary and sufficient conditions for this to be possible for two, three, and four dimensions.

3 Attainable convexity matrices

In this section we prove Theorem 3 which gives necessary and sufficient conditions for dimensions two, three, and four for when it is possible to create a multivariate Bernoulli $1/2$ with a given convexity matrix. We break the Theorem into three parts.

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 $B_1, B_2 \sim \text{Bern}(1/2)$ and $\mathbb{P}(B_1 = B_2) = \lambda_{12}$.

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:

$p_{10} + p_{11} = 0.5$, $p_{01} + p_{11} = 0.5$, $p_{11} + p_{00} = \lambda_{12}$, and $p_{00} + p_{01} + p_{10} + p_{11} = 1$. 

7
This system of linear equations has full rank, so there exists a unique solution. Given there exists a unique solution, it is easy to verify that

\[ p_{00} = (1/2)\lambda_{12}, \ p_{01} = (1/2)(1 - \lambda_{12}), \ p_{10} = (1/2)(1 - \lambda_{12}), \ p_{11} = (1/2)\lambda_{12} \]

satisfies the equations.

This provides an alternate algorithm to that found in [5] for simulating from bivariate distributions with correlation between \( \rho_{-12}, \rho_{+12} \).

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

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

Proof. Letting \( p_{ijk} = \mathbb{P}(B_1 = i, B_2 = j, B_3 = k) \), there are three conditions from the marginals.

\[ \sum_{j,k \in \{0,1\}} p_{1jk} = 0.5, \quad \sum_{i,k \in \{0,1\}} p_{i1k} = 0.5, \quad \sum_{i,j \in \{0,1\}} p_{ij1} = 0.5. \]

Then there are three conditions from the correlations

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

A seventh condition is \( \sum_{i,j,k} p_{ijk} = 1 \). Since eight equations are needed, suppose that \( p_{111} = \alpha \).

This 8 by 8 system of equations has full rank, so there is a unique solution. It is easy to verify that the solution is

\[ p_{000} = (1/2)(\lambda_{12} + \lambda_{13} + \lambda_{23} - 1) - \alpha \\
p_{001} = (1/2)(1 - (\lambda_{13} + \lambda_{23})) + \alpha \\
p_{010} = (1/2)(1 - (\lambda_{12} + \lambda_{23})) + \alpha \\
p_{011} = (1/2)\lambda_{23} - \alpha \\
p_{100} = (1/2)(1 - (\lambda_{12} + \lambda_{13})) + \alpha \\
p_{101} = (1/2)\lambda_{13} - \alpha \\
p_{110} = (1/2)\lambda_{12} - \alpha \\
p_{111} = \alpha \]
In order for this solution to yield probabilities, all must lie in \([0, 1]\). The \(p_{011}, p_{101},\) and \(p_{110}\) equations give the restriction

\[ \alpha \leq (1/2) \min\{\lambda_{12}, \lambda_{23}, \lambda_{13}\}. \]  

(2)

Of course, \(\alpha \geq 0\) so the \(p_{111}\) equation can be satisfied.

The \(p_{000}\) equation requires that

\[ \alpha \leq (1/2)(\lambda_{13} + \lambda_{12} + \lambda_{23} - 1). \]  

(3)

With these two conditions, equations \(p_{001}, p_{010},\) and \(p_{100}\) give the constraint

\[ \alpha \geq (1/2)(\lambda_{13} + \lambda_{12} + \lambda_{23} - \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\} - 1). \]  

(4)

The lower bound in (4) is below the upper bound in equation (3), but not necessarily below the upper bound in equation (2). So this gives another restriction:

\[ (1/2)(\lambda_{13} + \lambda_{12} + \lambda_{23} - \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\} - 1) \leq (1/2) \min\{\lambda_{13}, \lambda_{12}, \lambda_{23}\}. \]  

(5)

As long as (3) and (5) are satisfied, their will exist a solution. If satisfied with strict inequality, there will be an infinite number of solutions. \(\square\)

Note that not all positive definite correlation matrices are attainable with \(\text{Bern}(1/2)\) marginals. For instance, if \(\lambda_{12} = \lambda_{13} = \lambda_{23} = 0.3\), then \(\rho_{12} = \rho_{13} = \rho_{23} = -0.4\). With diagonal entries 1, the \(\rho\) values form a positive definite matrix, but it is impossible to build a multivariate distribution with \(\text{Bern}(1/2)\) marginals with these correlations.

4 Asymmetric Bernoulli distributions

Now consider the problem of drawing a multivariate Bernoulli \((X_1, \ldots, X_n)\) where \(X_i \sim \text{Bern}(p_i)\) where \(i\) is not necessarily 1/2, and the concurrence matrix is given.

**Lemma 3.** An \(n\) dimensional multivariate Bernoulli distribution where the marginal of component \(i\) is \(\text{Bern}(p_i)\) and concurrence matrix \(\Lambda\) exists if and only if an \(n + 1\) dimensional multivariate Bernoulli distribution exists with \(\text{Bern}(1/2)\) marginals and concurrence matrix

\[
\begin{pmatrix}
\Lambda & p_1^T \\
p_1 \cdot \cdots \cdot p_n & 1
\end{pmatrix}
\]
Proof. Suppose such an \( n + 1 \) dimensional distribution exists with Bern(1/2) marginals and specified concurrence matrix. Let \( (B_1, \ldots, B_{n+1}) \) be a draw from this distribution. Then set \( X_i = 1(B_i = B_{n+1}) \). The concurrence matrix gives \( P(X_i = 1) = p_i \), and for \( i \neq j \), \( P(X_i = X_j) = P(B_i = B_j) = \lambda_{ij} \).

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

\[
P(B_i = B_j) = P(X_i = X_j) = \lambda_{ij}.
\]

This result gives an alternate way of deriving Lemma 2 using the Fréchet-Hoeffding bounds. Consider a bivariate asymmetric Bernoulli \((X, Y)\) where \( X \sim \text{Bern}(p) \), and \( Y \sim \text{Bern}(q) \). The pseudoinverse cdf of \( \text{Bern}(p) \) is given by \( F_p^{-1}(U) = 1(U > 1 - p) \). Hence the minimum and maximum correlation of \( X \) and \( Y \) can be directly found using Theorem 1:

\[
\rho^-(X, Y) = \frac{(p + q - 1)(p + q > 1) - pq}{\sqrt{pq(1-p)(1-q)}}
\]

\[
\rho^+(X, Y) = \frac{\min(p, q) - pq}{\sqrt{pq(1-p)(1-q)}}
\]

Suppose that we want to simulate \( X \) and \( Y \) so that \( P(X = Y) = r \). For what \( r \) is this possible?

Let \( p_{ij} = P(X = i, Y = j) \). Then

\[
P(X = Y) = r = p_{11} + p_{00}
\]

\[
P(X = 1) = p = p_{10} + p_{11}
\]

\[
P(Y = 1) = q = p_{01} + p_{11}
\]

Adding the three equations above we have

\[
p + q + r = p_{11} + p_{00} + p_{10} + p_{11} + p_{01} + p_{11} = 1 + 2p_{11}
\]

and therefore \( p_{11} = (p + q + r - 1)/2 \). On the other hand

\[
E(XY) = E(1(X = 1, Y = 1)) = P(X = 1, Y = 1) = p_{11}.
\]

The Fréchet-Hoeffding Theorem provides the condition

\[
E(F_p^{-1}(U)F_q^{-1}(1 - U)) \leq E(XY) \leq E(F_p^{-1}(U)F_q^{-1}(U))
\]

10
which is equivalent to
\[(p + q - 1)1(p + q > 1) \leq (p + q + r - 1)/2 \leq \min(p, q).
\]

Solving the inequalities for \(r\) we get the following condition
\[|1 - (p + q)| \leq r \leq 2\min(p, q) + 1 - (p + q).
\]

Letting \(p = \lambda_{13},\ q = \lambda_{23}\), and \(r = \lambda_{12}\) we obtain an equivalent condition to the one in Lemma 2.

As a further consequence of this new proof technique, we obtain a simple method for simulating from the 3 dimensional symmetric Bernoulli distribution for given concurrence probabilities when possible.

<table>
<thead>
<tr>
<th>Generating trivariate symmetric Bernoulli</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> 3 by 3 matrix ((\lambda_{ij})) with (1 \leq \lambda_{12} + \lambda_{13} + \lambda_{23} \leq 1 + 2 \min{\lambda_{12}, \lambda_{13}, \lambda_{23}}).</td>
</tr>
<tr>
<td><strong>Output:</strong> ((B_1, B_2, B_3)) where ((\forall i)(B_i \sim \text{Bern}(1/2))) and ((\forall i, j)(\mathbb{P}(B_i = B_j) = \lambda_{ij}))</td>
</tr>
<tr>
<td>1) Draw (B_3 \leftarrow \text{Bern}(1/2)) and (U \leftarrow \text{Unif}([0, 1])).</td>
</tr>
<tr>
<td>2) (X_1 \leftarrow 1(U \in [0, \lambda_{13}]))</td>
</tr>
<tr>
<td>3) (X_2 \leftarrow 1(U \in [(1 + \lambda_{13} - \lambda_{23} - \lambda_{12})/2, (1 + \lambda_{13} + \lambda_{23} - \lambda_{12})/2]))</td>
</tr>
<tr>
<td>4) (B_1 \leftarrow B_3X_1 + (1 - B_3)(1 - X_1))</td>
</tr>
<tr>
<td>5) (B_1 \leftarrow B_3X_2 + (1 - B_3)(1 - X_2))</td>
</tr>
</tbody>
</table>

**Proposition 2.** The above algorithm simulates \((B_1, B_2, B_3)\) where \(B_i \sim \text{Bern}(1/2)\) for all \(i\) and \(\mathbb{P}(B_i = B_j) = \lambda_{ij}\) for all \(i\) and \(j\).

**Proof.** Note \(1 \leq \lambda_{12} + \lambda_{13} + \lambda_{23} \leq 1 + 2 \min\{\lambda_{12}, \lambda_{13}, \lambda_{23}\}\) if and only if
\[0 \leq \frac{1 + \lambda_{13} - \lambda_{23} - \lambda_{12}}{2} \leq \lambda_{13} \leq \frac{1 + \lambda_{13} + \lambda_{23} - \lambda_{12}}{2} \leq 1.
\]

Hence
\[
\mathbb{P}(X_1 = 1) = \lambda_{13} - 0 = \lambda_{13},
\]
\[
\mathbb{P}(X_2 = 1) = \frac{1 + \lambda_{13} + \lambda_{23} - \lambda_{12}}{2} - \frac{1 + \lambda_{13} - \lambda_{23} - \lambda_{12}}{2} = \lambda_{23}.
\]
\[
\mathbb{P}(X_1 = X_2) = \left(\frac{\lambda_{13} - \frac{1 + \lambda_{13} - \lambda_{23} - \lambda_{12}}{2}}{2}\right) + \left(1 - \frac{1 + \lambda_{13} + \lambda_{23} - \lambda_{12}}{2}\right) = \lambda_{12}.
\]

Then as in the proof of Lemma 3, the output has the desired trivariate symmetric Bernoulli distribution. \qed
Lemma 3 also allows a characterization of the 4 dimensional symmetric Bernoulli concurrence (and so also convexity) matrices.

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

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

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

it is true that

\[
1 - (1/2)\ell \leq (1/2)(\nu - 1)
\]

Proof. By using Lemma 3, the problem is reduced to finding a distribution for \((X_1, X_2, X_3)\) where \(X_i \sim \text{Bern}(\lambda_{ii})\) and the upper 3 by 3 minor of \(\Lambda\) is the new concurrence matrix. Just as in Lemma 2, this gives eight equations of full rank with a single parameter \(\alpha\). Letting \(q_{ijk} = P(X_1 = i, X_2 = j, X_3 = k)\), the unique solution is

\[
q_{000} = (1/2)(\lambda_{12} + \lambda_{13} + \lambda_{23}) - (1/2) - \alpha
\]

\[
q_{001} = -(1/2)(\lambda_{14} + \lambda_{24} + \lambda_{13} + \lambda_{23}) + 1 + \alpha
\]

\[
q_{010} = -(1/2)(\lambda_{14} + \lambda_{34} + \lambda_{12} + \lambda_{23}) + 1 + \alpha
\]

\[
q_{011} = (1/2)(\lambda_{24} + \lambda_{34} + \lambda_{13}) - (1/2) - \alpha
\]

\[
q_{100} = -(1/2)(\lambda_{24} + \lambda_{34} + \lambda_{12} + \lambda_{13}) + 1 + \alpha
\]

\[
q_{101} = (1/2)(\lambda_{14} + \lambda_{34} + \lambda_{13}) - (1/2) - \alpha
\]

\[
q_{110} = (1/2)(\lambda_{34} + \lambda_{24} + \lambda_{12}) - (1/2) - \alpha
\]

\[
q_{111} = \alpha.
\]

All of these right hand sides lie in \([0,1]\) if and only if \(1 - (1/2)\ell \leq (1/2)(\nu - 1)\), and \(\alpha\) is chosen to lie in \([1 - (1/2)\ell, (1/2)(\nu - 1)]\).

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

5 Conclusions

The Fréchet-Hoeffding bounds give a lower and upper bound on the pairwise correlation between two random variables with given marginals. Hence
for higher dimensions the correlation matrix can be written as a convexity matrix whose parameters indicated where on the line from the lower to the upper bound the correlation lies. With a simple algorithm, we have shown that when viewed as a convexity matrix problem, the worse case for marginals is when they are all symmetric Bernoulli random variables. It is worst case in the sense that if it is possible to build a symmetric Bernoulli multivariate distribution with the given convexity matrix, then it is possible to build a multivariate distribution with that convexity matrix for any marginal distributions. For symmetric Bernoulli marginals, the convexity matrix is also the concurrence matrix that gives the probabilities that any pair of random variables are the same.

For 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


