Generating random correlation matrices with constraints

by

Isabel van der Brug

to obtain the degree of Bachelor of Science
at the Delft University of Technology,
to be defended publicly on Friday 4th July.

Student number: 5697069
Project duration: April 21, 2025 –June 27 ,2025
Thesis committee: D. Kurowicka, TU Delft, supervisor
N. Parolya, TU Delft

TU Delft

Lay summary

In many areas of science and finance it can be useful to understand how different factors relate to each other, whether they move together or independently. A way to express these relationships is through a correlation matrix, where each entry shows how strongly a pair of variables is related. These matrices are widely used to make predictions, for example, in understanding how different assets behave in financial markets. However, what if not all of the data is available and we still want to make predictions or test models? In such cases we need to simulate correlation matrices that are both realistic and mathematically valid. This thesis explores two techniques for generating these matrices, along with extensions that allow the average correlation to be fixed or controlled. The first method is based on square root decomposition, which builds the matrix step by step from a set of unit vectors. An extension of this method allows us to fix the average correlation which is useful in applications like financial risk management, where we may want to model scenarios with a known level of average dependence between assets. The second method is based on probabilistic principles, rather than geometric construction. This gives us more control over the statistical properties of the matrices. One key extension is the ability to fix the expected correlation, making this method ideal for simulations where we want to influence the average behaviour without enforcing it exactly. By comparing these two methods, this thesis provides deeper insight into how they work, how they differ, and when one is more suitable. This helps guide the choice of the right method for generating synthetic but realistic data structures in different applications.

Summary

Correlation matrices play a central role in multivariate modelling across fields such as finance and statistics. However, generating valid correlation matrices, remains a non-trivial problem due to the global positive definiteness condition they must satisfy. This thesis investigates two methods for generating correlation matrices, with extensions on how to control or influence the average correlation.

The first method relies on square root decomposition of the correlation matrix, parametrizing it as the product of a matrix with unit-norm rows and its transpose. A recent extension of this by Tuitman et al. [14] is explored, which enables the generation of matrices with a fixed average correlation. This is achieved through iterative construction of the decomposition, ensuring the weighted sum of vectors has a prescribed norm, corresponding to the target average correlation. The algorithms geometric structure, feasibility conditions, and statistical properties are analysed.

The second method is based on the C-vine construction using partial correlations, as introduced by Joe and Kurowicka [10]. There exists a one-to-one mapping from a set of partial correlations to a full correlation matrix. This approach parametrizes the matrix through a structured sequence of partial correlations. The distribution from which these partial correlations are sampled can be adjusted to achieve specific properties in the resulting matrices, for example using specific Beta distributions we obtain matrices following the LKJ-distribution. The extension by Joe and Kurowicka [10] is investigated, which allows the expected value of each correlation to be fixed across samples.

A comparison of both methods is provided in terms of construction, flexibility, numerical stability, and statistical properties of the resulting matrices. While the square root decomposition method offers strict per-matrix control over the average correlation, the C-vine approach provides greater flexibility, enabling finer control over marginal distributions. The thesis concludes with a discussion on practical trade-offs and potential directions for future work.

All of the figures and data presented in this thesis was computed in R-studio. For access to the implementations or underlying code.

Contents

1 Introduction 4

2 Generating correlation matrices using square root decomposition parametrization 7
- 2.1 Unconstrained square root decomposition parametrization . . . 8
  - 2.1.1 Example . . . 8
  - 2.1.2 Properties of the matrices generated using SRD parametrization . . . 8
- 2.2 Generating correlation matrices with SRD parametrization and average correlation constraint . . 10
  - 2.2.1 Geometric intuition behind the algorithm . . . 10
  - 2.2.2 Theoretical background . . . 11
  - 2.2.3 Implementation and results . . . 16
  - 2.2.4 Example . . . 17
  - 2.2.5 Properties of the matrices generated with the algorithm by Tuitman et al . . . 19
  - 2.2.6 Comparison of the results for matrices generated with and without average constraint by SRD parametrization . . . 24

3 Generating correlation matrices using partial correlation parametrization 25
- 3.1 Properties of the C-vine Partial correlation parametrization . . . 30
  - 3.1.1 Scatter plots and marginal distributions of sampled correlation matrices with LKJ distribution . . . 30
  - 3.1.2 Example using asymmetric Beta partial correlations . . . 32
  - 3.1.3 Permutation-based symmetrization of correlation matrix entries . . . 33
- 3.2 Fixing the expectation of each correlation . . . 34
  - 3.2.1 Empirical properties of correlation matrices fixed expectations . . . 35
  - 3.2.2 Distribution of average correlation . . . 39

4 Summary and discussion 41
- 4.1 Theoretical foundations and constraints . . . 41
- 4.2 Statistical properties of the matrices . . . 42
- 4.3 Numerical Stability, Implementation, and Control . . . 42
- 4.4 Discussion . . . 43

Bibliography 45

A Appendix A 47

1

Introduction

Correlation matrices $C = [C_{ij}]_{i,j=1,\dots,n}$ are $n \times n$ symmetric matrices with ones on the diagonal and off-diagonal elements in the interval $[-1,1]$. These matrices are also positive semi-definite, hence all eigenvalues are non-negative. The positive semi-definiteness imposes a global constraint, not all combinations of off-diagonal elements on $[-1,1]$ leads to a valid matrix. As a result, the set of all correlation matrices form a subset of the hypercube $[-1,1]^{\frac{n(n-1)}{2}}$.

Correlation matrices are widely applied in probability and statistics to help visualize relationships, identify similarities and patterns within a data set. They can serve as an input to statistical models and methods such as regression and factor analysis. In simulation-based methods, such as Monte Carlo analysis, it is often necessary to generate random correlation matrices to study the behaviour of complex multivariate systems [7]. This approach involves repeatedly simulating correlation structures to estimate quantities such as the expected value, variance, and distributional properties of system outputs, thereby allowing for statistical inference and pattern prediction under uncertainty. Different methods to generate correlation matrices can be applied for these simulations, each with various properties and aims. A simple method to generate random correlation matrices without underlying data is the construction proposed by Marsaglia and Olkin [12], who found that C is a correlation matrix if and only if there exists T such that $C = TT'$, where T' is the transpose and the rows of T must be unit vectors. In this construction, the correlation C_{ij} corresponds to the inner product of the i-th and j-th row of T. Hence we can generate random correlation matrices by sampling rows of T from the unit sphere.

In many practical situations, only partial information about the correlation structure is available, for instance only some pairwise correlations are given. In such cases, a class of recursive algorithms can be used to complete the matrix while preserving positive semi-definiteness. As discussed in the comparative study by Flórez et al. [6], these methods construct the matrix by determining the feasible interval for each new correlation entry based on the already specified values. These techniques are relevant in clinical research for example, these matrices can allow researchers to explore the possible dependency structures and how strongly a surrogate (such as laboratory result) predicts real health outcomes.

Another application is proposed by Hüttner and Mai [7], where correlation matrices that satisfy the perron frobenius property are simulated. Matrices with this property possess the dominant eigenvector with strictly positive entries. The positivity of the dominant eigenvector is important as the entries of this vector are used to approximate the optimal choice of portfolio weights for correlated stocks[4].

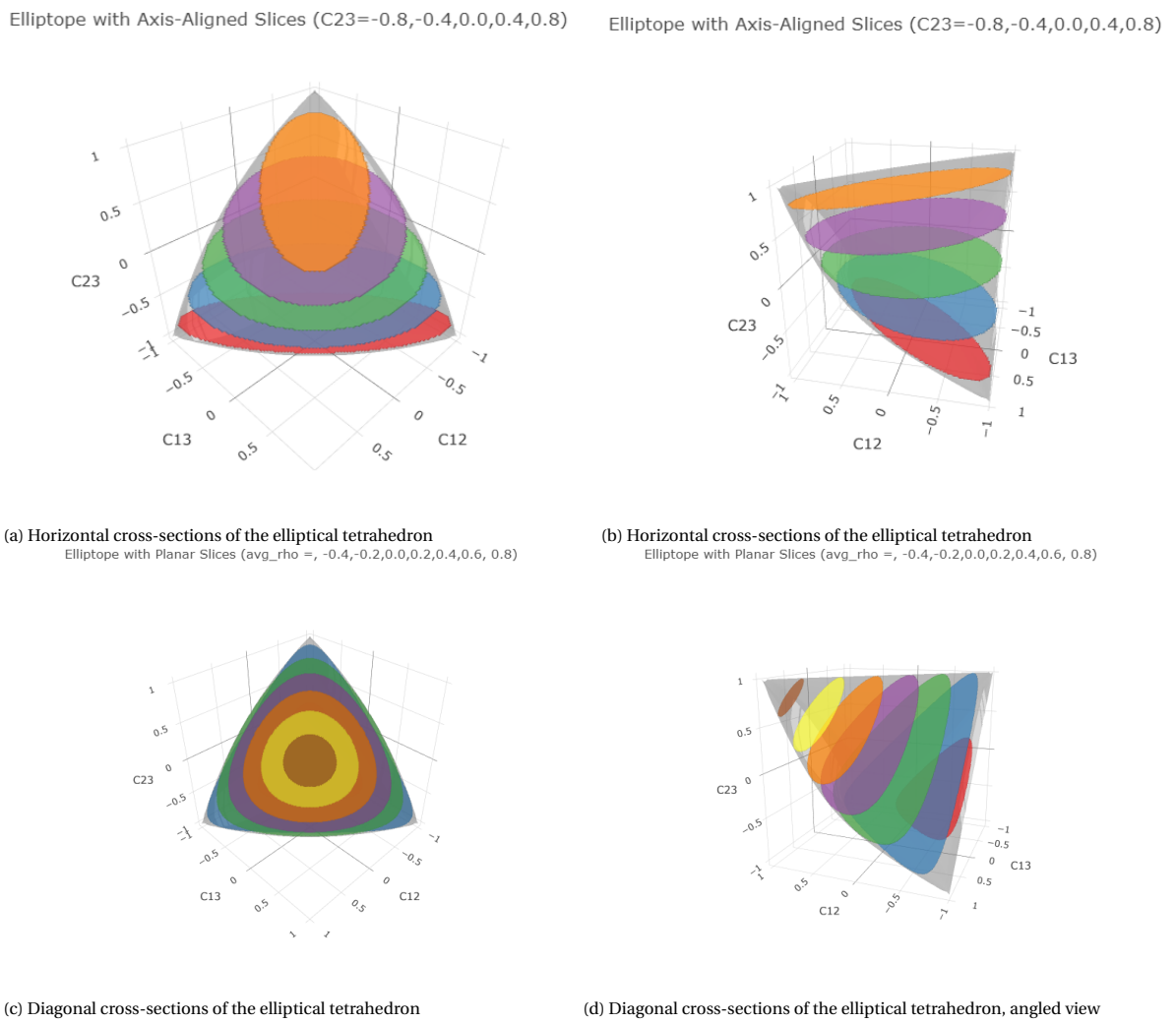
In addition, methods exist for generating correlation matrices uniformly from a known distribution. A notable example is the LKJ distribution, introduced by Lewandowski et al. [11], which defines a family of distributions over the space of correlation matrices. In this approach, the joint density of the correlations in the matrix is proportional to a power of the determinant of the matrix. This allows one to control how strongly the samples are concentrated around the identity matrix. The LKJ distribution is particularly well suited for simulation and Bayesian modelling, where structured priors over correlation matrices are often required.

In all of these methods, the main challenge is to ensure that the resulting matrices remain positive semi-

definite while satisfying additional constraints. In high dimensions, the space of valid correlation matrices is difficult to visualize, it becomes much more tractable in the low dimensional $n = 3$ case. By plotting the off-diagonal elements C_{12}, C_{13} and C_{23} of valid 3×3 correlation matrices, one can visualize points in the cube $[-1, 1]^3$. However, due to the positive semi-definite constraint, the admissible set forms a convex subset of this cube. This region is symmetric under permutations of C_{12}, C_{13} and C_{23}. [5] Every point within this space satisfies the determinant condition:

$$\det(C) = 1 + 2C_{12}C_{13}C_{23} - C_{12}^2 - C_{13}^2 - C_{23}^2 \geq 0 \tag{1.1}$$

The set of valid correlation matrices, denoted S_3 is presented in Figure 1.1.



(a) Horizontal cross-sections of the elliptical tetrahedron

(b) Horizontal cross-sections of the elliptical tetrahedron

(c) Diagonal cross-sections of the elliptical tetrahedron

(d) Diagonal cross-sections of the elliptical tetrahedron, angled view

Figure 1.1: Figures of elliptical tetrahedron for n=3 with horizontal and vertical cross-sections.

The set S_3 is known as an elliptical tetrahedron or elliptope. It is convex body with 4 sharp vertices connected by 6 curved edges. The surface is smooth along these edges. At the intersections with the planes $C_{12} = 0, C_{13} = 0$ and $C_{23} = 0$, the elliptical tetrahedron intersects three orthogonal circles can be found [5].

Figure 1.1a shows that the cross-sections parallel to a coordinate plane are ellipses (here the case when $C_{23} = b$ with $|b| \leq 1$ is shown). This is the situation we mentioned earlier, where only partial information is available. When $b = 0$, the cross section is a circle. For positive b the major axis of the ellipse is in the direction of $C_{12} = C_{13}$, and minor axis in direction $C_{12} = -C_{13}$, these roles reverse when $b < 0$[13]. This axis swap is clearly visible in Figures 1.1a and 1.1b, where the direction of the major changes as b passes through zero.

Figure 1.1c shows intersections between the elliptical tetrahedron with the planes $C_{12} + C_{13} + C_{23} = d$, for $|d| \leq 3$. The cross-sections appear as rounded triangles when $d < 0$, gradually becoming more circular as

d increases. These planar slices correspond to sets of correlations with fixed average correlation, since the plane $C_{12} + C_{13} + C_{23} = d$ imposes the condition that all matrices have average correlation $\frac{d}{3}$.

There exist different methods to find the matrices that satisfy an average correlation ρ. This can be done by simply sampling all possible matrices with off-diagonal values between $[-1,1]$ that satisfy the average constraint and rejecting all matrices that are not positive semi-definite. However, Tuitman et al. [14] present an efficient method to generate correlation matrices that lie on the hyperplane where average off-diagonal correlation equals to a prescribed target value. Each such matrix can be considered as a parameter of a multivariate normal distribution of $(X_1, ..., X_n)$, where the average correlation directly determines the variance of the sum $\sum X_i$. This theory can be applied in risk aggregation as measures like Value-at-Risk depend on total variance [3] and help to study extreme dependence scenarios under Gaussian marginals Wang et al. [15]. There also exist more probabilistic approaches such as the method by Joe and Kurowicka [10] which extends partial correlation c-vine parametrization so enforce a fixed expectation for marginals. This method can be applied in Bayesian modelling to specify prior distributions whose average structure matches prior beliefs.

Different applications require different constraints on correlation matrices, which result in distinct geometric structures. This thesis focuses on two approaches for generating valid correlation matrices, square root decomposition parametrization and the extension of this by Tuitman et al. [14], and partial correlation C-vine parametrization and extension by Joe and Kurowicka [10]. These methods operate within the space of valid correlation matrices in different ways. Tuitman et al. [14] introduces an iterative procedure to generate correlation matrices with a given average constraint, making it suitable for risk aggregation contexts. In contrast, using the partial correlation parametrization, we can sample matrices with prescribed expected value of each correlation. Therefore, matrices for which the expected average correlation is specified can also be simulated.

The goal of this thesis is to present these different methods, evaluate similarities and differences and present possible extensions.

The outline of the thesis is as follows: In Section 2 the method of generation the correlation matrices based of square root decomposition is presented. This method is then extended following Tuitman et al. [14] to simulate matrices with specified average correlation. We examine the construction by presenting geometric and probabilistic properties of the algorithm. In Section 3 the partial correlation parametrization method is presented, which is then extended to fix the expectation of each correlation. In Section 4 a comparison between the two methods is presented and possible extensions left as a future work are proposed.

2

Generating correlation matrices using square root decomposition parametrization

In this section, we investigate the method of generating correlation matrices using thesquare root decomposition parametrization, where the rows of T are unit vectors. We then introduce an additional constraint of fixing the average correlation, as proposed by Tuitman et al. [14], and examine how this alters the construction. An algorithm for generating such matrices is presented, followed by an exploration of the geometric properties of this space. Throughout this chapter square root decomposition will be referred to as SRD. We begin by introducing notation that will be used throughout this chapter.

Table 2.1: Summary of notation used in the algorithm by Tuitman et al.

Symbol	Meaning
n	Dimension of the correlation matrix ($n \times n$)
σ_i	user-defined weights assigned influence to each correlation coefficient
C	Correlation matrix to be generated, C_{ij} is matrix element from row i and column j
T	Matrix in $\mathbb{R}^{n\times n}$ such that $C = TT^\top$, with unit-norm rows
T'	The transpose of matrix T
N_j	Standard normal random variable
X_i	Random variable defined by $\sigma_i \sum_{j=1}^{n} T_{ij} N_j$
s^2	The variance of the sum of random variables X_i
t_i	i-th row of T, a unit vector in $\mathbb{R}^n$
u_i	Partial weighted sum: $u_i = \sum_{j=1}^{i} \sigma_j t_j$
l_i	Norm of partial sum: $l_i = \|u_i\|$
s	Target norm of the full weighted sum: $s = \left\|\sum_{i=1}^{n} \sigma_i t_i\right\|$
ρ	Desired average pairwise correlation in the matrix C
$\langle t_i, t_j \rangle$	Inner product of t_i and t_j, which is also C_{ij}
$\|\cdot\|$	the euclidean norm

A key result from Marsaglia and Olkin [12] demonstrates that a matrix is positive definite if it can be rewritten as a product of two other matrices.

Theorem 1. *Let C be a n by n symmetric matrix with ones on the main diagonal.C is positive definite if and only if there exists $T \in M_{nxn}(\mathbb{R})$ such that $C = TT'$ and the rows of T must be vectors in $\mathbb{R}^n$ of length 1.*

The inner product of row i and j of T hence give element C_{ij}. The decomposition above is unique if the matrix T is a lower triangular matrix.

2.1. Unconstrained square root decomposition parametrization

We can sample unit vectors t_i uniformly and independently from the unit sphere and placed them as rows of the matrix T. The correlation matrix C is then obtained by computing the matrix product TT'. The algorithm is presented below.

Algorithm 1 Generation of Random Correlation Matrix via SRD parametrization

1: **Given:** $n \in \mathbb{N}$
2: **Generate:** Correlation matrix $C \in \mathbb{R}^{n\times n}$
3: Initialize $T \leftarrow 0 \in \mathbb{R}^{n\times n}$
4: **for** $i := 1$ to n **do**
5: $\quad t_i \leftarrow$ random in $\mathbb{R}^n$ of length 1
6: $\quad$ Set $T[i,:] \leftarrow \mathbf{t}_i$
7: **end for**
8: Compute $C \leftarrow TT'$
9: **Return:** C

2.1.1. Example

A step by step application of the algorithm for $n = 3$ is presented. The vectors are sampled uniformly from (-1,1) and normalized:

$$t_1 = \begin{bmatrix} -0.223 \\ -0.816 \\ 0.533 \end{bmatrix}, \quad t_2 = \begin{bmatrix} 0.515 \\ 0.811 \\ 0.276 \end{bmatrix}, \quad t_3 = \begin{bmatrix} -0.218 \\ 0.816 \\ 0.535 \end{bmatrix}$$

which yields the matrix:

$$T = \begin{bmatrix} -0.223 & -0.816 & 0.533 \\ 0.515 & 0.811 & 0.276 \\ -0.218 & 0.816 & 0.535 \end{bmatrix}$$

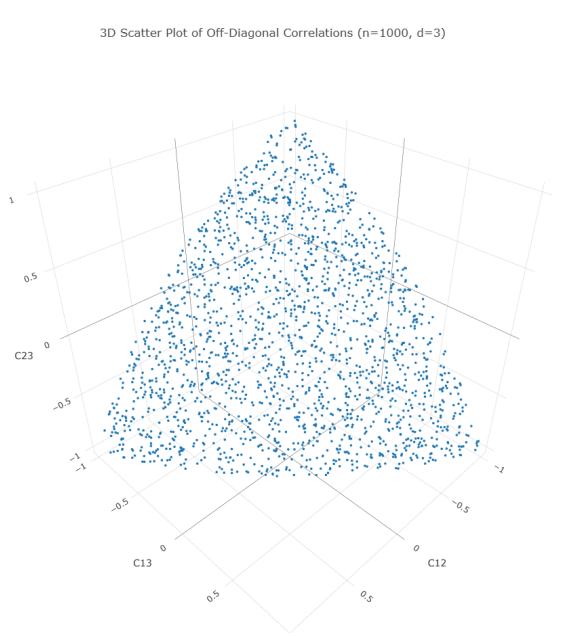
The corresponding correlation matrix $C = TT'$ is then:

$$C = \begin{bmatrix} 1 & -0.630 & -0.333 \\ -0.630 & 1 & 0.698 \\ -0.333 & 0.0.698 & 1 \end{bmatrix}$$

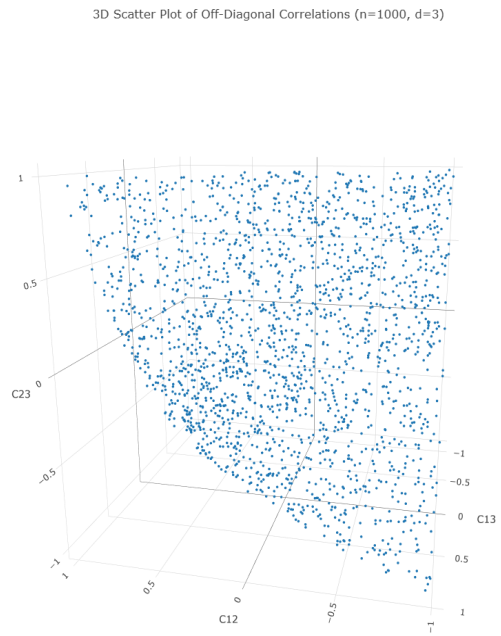
2.1.2. Properties of the matrices generated using SRD parametrization

This subsection examines the behaviour of correlation matrices generated using Algorithm 1. We analyse both the marginal distributions of the off-diagonal entries and the distribution of the matrices within the feasible region.

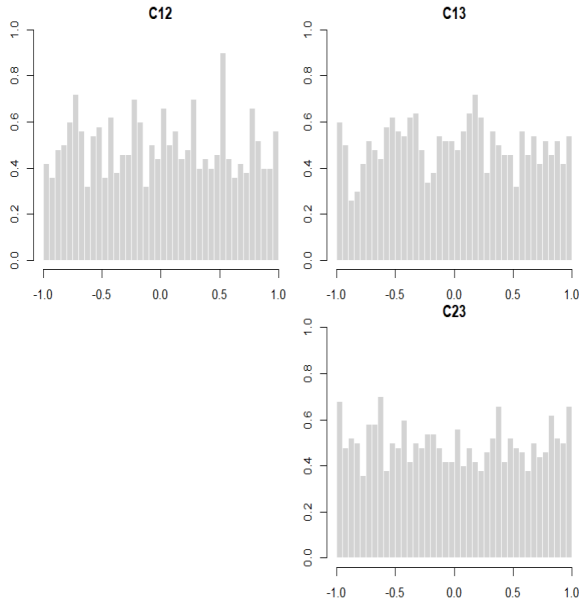
Figure 2.1 confirms that the generated matrices indeed lie within the feasible region of 3×3 correlation matrices. Furthermore, they appear to be uniformly distributed. This is expected, given that the vectors t_i are sampled uniformly from the unit sphere, and the resulting inner products $C_{ij} = \langle t_i, t_j \rangle$ lie in the elliptical tetrahedron. This is further supported by the behaviour marginal distributions in Figure 2.1c.



(a) 1000 randomly generated correlation matrices



(b) 1000 randomly generated correlation matrices



(c) Marginal distribution of C_{ij}

Figure 2.1: 3D scatter plots and marginal distribution for correlation matrices generated using SRD parametrization.

In Table 2.2, results regarding the empirical means, variances and pairwise correlations over 1000 samples are summarized. The empirical means are all very close to zero. This is consistent with the fact that the rows of matrix T are independent vectors sampled from the unit sphere, resulting in correlations that are symmetrically distributed around zero. The small fluctuations can be attributed to the fact that we are using a finite sample (N=1000).
The variances of the off-diagonal elements C_{ij} in Table 2.2 are all observed to be approximately $\frac{1}{3}$, this result can be derived as follows:
The mean of each $C_{ij} = 0$, and the variance simplifies to the second moment:

$$\mathrm{Var}(C_{ij}) = \mathbb{E}(C_{ij}^2) - \mathbb{E}(C_{ij})^2 = \mathbb{E}(\langle t_i, t_j \rangle^2)$$

Expanding the dot product, we obtain:

$$\mathbb{E}\left(\left(\sum_{k=1}^{n} t_{ik}t_{jk}\right)^2\right) = \sum_{k=1}^{n}\sum_{\ell=1}^{n}\mathbb{E}(t_{ik}t_{i\ell})\mathbb{E}(t_{jk}t_{j\ell})$$

Due to independence and spherical symmetry, all terms $\mathbb{E}(t_{ik}t_{il})$ where $k \neq \ell$ are equal to zero, and the terms $\mathbb{E}(t_{ik}^2) = \frac{1}{n}$. Thus we obtain:

$$\sum_{k=1}^{n}\mathbb{E}(t_{ik}^2)\mathbb{E}(t_{jk}^2) = n\left(\frac{1}{n}\cdot\frac{1}{n}\right) = \frac{1}{n}$$

This result holds for all n.

The pairwise correlation between off-diagonal elements are all near zero, indicating that these elements are uncorrelated with each other. This is expected, as each row of T is sampled independently, and off-diagonal entries are functions of independent unit vectors. Therefore, although the matrix is subject to the positive semi-definite constraint, this does not introduce strong dependencies between off-diagonal elements.

Table 2.2: Simulation summary of 1000 matrices

Expected Correlation			**Variance**			**Correlation between Elements**		
C_{12}	C_{13}	C_{23}	C_{12}	C_{13}	C_{23}	$C_{12,13}$	$C_{12,23}$	$C_{13,23}$
0.004	-0.002	-0.008	0.337	0.317	0.338	0.009	-0.001	0.009

2.2. Generating correlation matrices with SRD parametrization and average correlation constraint

Construction by $C = TT'$ offers flexibility, as specific properties of the correlation matrix C can be controlled by carefully designing the structure of the rows of T. One example is the algorithm proposed by Tuitman et al. [14], where the rows of T are constructed in such a way that the resulting matrix C has a prescribed average correlation. This section will provide a detailed analysis of this algorithm. We begin by providing geometric intuition behind the construction, then the method is formally presented. Next the implementation of the method is presented and an example is show for $n = 3$. Finally, properties of generated with this method matrices are studied.

2.2.1. Geometric intuition behind the algorithm

In order to impose an additional constraint, the construction of T in the square root decomposition requires careful design. The (i, j)th entry of correlation matrix C is determined by the inner product of the i and jth rows, the respective orientation of the vectors must be controlled. Tuitman et al. [14] propose a method to construct these vectors such that the angles between them satisfy desired constraints. Indirectly an average correlation can be achieved by controlling the length of the weighted sum of the rows of T. Geometrically this corresponds to requiring the weighted sum of the vector to lie exactly on the surface of a sphere with a fixed radius. The radius of the sphere corresponds to the average correlation: if the vectors are closely aligned then the angles between them will be smaller and correlations in the matrix will be stronger. In this case the result is that the sum of the vectors t_i is longer. If the vectors are not aligned, the sum is shorter and the correlation values are weaker. This can be interpreted as incrementally constructing the matrix within the feasible region of valid correlation matrices, where each step is carefully constrained so that the final matrix lies exactly on the hyperplane corresponding to the desired average correlation. Figure 2.2 shows an example of the process of building up these vectors t_i for $n = 3$.

The vectors t_i must be so that successive vectors are not collinear, and attain the correct length. Hence t_i is constructed with two components. A component in the direction of the sum $\sum_{j=1}^{i-1} t_j$ is defined to control the growth of the cumulative sum, ensuring the feasibility of reaching final norm s. The length of this new vector lies between 1 and s. In addition, an orthogonal component to this direction prevents the vectors t_i from being collinear. Therefore, at each step the orientation and length are adjusted within precise bounds.

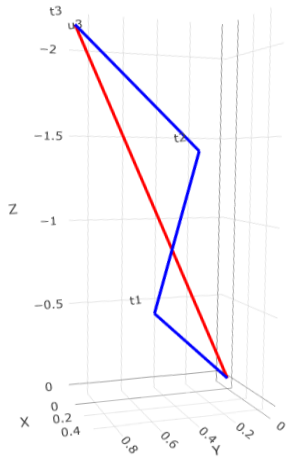


Figure 2.2: plot of the vectors t_1, t_2 and t_3 in blue for $\rho = 0.5$ and the resulting vector u_3 in red with length s

The connection between s and ρ can be demonstrated in the 3x3 case:

$$\frac{C_{12} + C_{13} + C_{23}}{3} = \frac{\langle t_1, t_2\rangle + \langle t_1, t_2\rangle + \langle t_2, t_3\rangle}{3} = \rho$$

The way that this algorithm ensures this happens is by constructing t_i such that eventually $\| \sum_{i=1}^{n} t_i \|^2 = s^2$

$$\text{We aim for } \rho = \frac{\langle t_1, t_2\rangle + \langle t_1, t_3\rangle + \langle t_2, t_3\rangle}{3}$$

$$\begin{aligned}
\text{Now consider: } s^2 &= \| t_1 + t_2 + t_3 \|^2 \\
&= \langle t_1 + t_2 + t_3, t_1 + t_2 + t_3\rangle \\
&= \| t_1 \|^2 + \| t_2 \|^2 + \| t_3 \|^2 + 2\langle t_1, t_2\rangle + 2\langle t_1, t_3\rangle + 2\langle t_2, t_3\rangle \\
&= 3 + 2\left(\langle t_1, t_2\rangle + \langle t_1, t_3\rangle + \langle t_2, t_3\rangle\right) \\
&= 3 + 6\rho \\
&\Rightarrow \frac{\langle t_1, t_2\rangle + \langle t_1, t_3\rangle + \langle t_2, t_3\rangle}{3} = \rho
\end{aligned}$$

2.2.2. Theoretical background

This section provides a theoretical explanation of the algorithm proposed by Tuitman et al. [14]. We begin by reformulating the average correlation constraint in terms of the total variance of a weighted sum of Gaussian variables. This leads to the geometric interpretation of constructing unit-norm vectors whose weighted sum has a prescribed norm. We then present the necessary conditions and intervals that ensure the feasibility of the construction, and explain the iterative procedure used to build the matrix T row by row, ultimately producing the final correlation matrix $C = TT'$.

The theory is presented for the weighted sum of Gaussian variables but in this thesis only the uniform weights will be applied. The weights can be used when additional information about the quality of some correlations in the correlation matrix are given.

Objective:

The aim of the algorithm presented by Tuitman et al. [14] is to generate correlation matrices with a given average correlation value. This means that $C \in \mathbb{M}_{n\times n}(\mathbb{R})$ must be such that:

$$\frac{\sum_{i<j} \sigma_i \sigma_j C_{ij}}{\sum_{i<j} \sigma_i \sigma_j} = \rho \tag{2.1}$$

where $\sigma_i, i = 1, ..., n$ denote the weights.

The constraint can be reformulated in the following form:

$$\rho = \frac{s^2 - \sum_{i=1}^{n} \sigma_i^2}{2\sum_{i<j} \sigma_i \sigma_j}, \tag{2.2}$$

where s^2 is the total variance of a linear combination of random variables X_i (defined in equation 2.3), and the denominator is a fixed constant. This reformulation reveals that fixing the average correlation ρ is equivalent to fixing the total variance s^2 of a weighted sum of the X_i. The next step is to construct these variables and show that this constraints are indeed equivalent.

To make this connection precise, observe the identity:

$$\sum_{i,j=1}^{n} \sigma_i \sigma_j C_{ij} = \sum_{i=1}^{n} \sigma_i^2 + 2 \sum_{i<j} \sigma_i \sigma_j C_{ij},$$

A key insight is that $\sum_{i,j=1} \sigma_i \sigma_j C_{ij}$ is equal to the variance (s^2) of $\sum_i^n X_i$ for X_i's defined below.

$$X_i = \sigma_i \sum_{j=1}^{n} T_{ij} N_j, i = 1, \ldots, n \tag{2.3}$$

where N_j's are independent standard Gaussian random variables.

We start with the following lemma.

Lemma 1. *The total variance of the sum of random variables X_i is given by:*

$$s^2 = Var(\sum_{i=1}^{n} X_i) = \sum_{i,j=1}^{n} \sigma_i \sigma_j C_{ij}$$

Proof. We have

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i,j=1}^{n} \mathrm{Cov}(X_i, X_j).$$

Since

$$X_i = \sigma_i \sum_{k=1}^{n} T_{ik} N_k,$$

where $T \in \mathbb{R}^{n \times n}$ has rows t_i, and $N_k \sim \mathcal{N}(0,1)$ are independent standard normal variables. then

$$\mathbb{E}[N_j N_\ell] = \begin{cases} 1 & \text{if } j = \ell, \\ 0 & \text{if } j \neq \ell. \end{cases}$$

Using this, the covariance between X_i and X_j becomes:

$$\mathrm{Cov}(X_i, X_j) = \sigma_i \sigma_j \sum_{k=1}^{n} T_{ik} T_{jk} = \sigma_i \sigma_j \langle t_i, t_j \rangle.$$

Since $C = TT^\top$, this gives:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i,j=1}^{n} \sigma_i \sigma_j C_{ij}.$$

□

Corollary 1. *By lemma 1 we can conclude that this problem can also be interpreted as generating matrices that satisfy the variance constraint.*

$$s^2 = \| \sigma_1 t_1 + \sigma_2 t_2 + \cdots + \sigma_n t_n \|^2 \tag{2.4}$$

Finally, if $\sum_{i,j=1}^{n} \sigma_i \sigma_j C_{ij} = s^2$, then we can conclude using these results that average correlation ρ satisfies:

$$\rho = \frac{s^2 - \sum_{i=1}^{n} \sigma_i^2}{2\sum_{i<j} \sigma_i \sigma_j}. \tag{2.5}$$

If $\sigma_i = 1$ for all i , then we have :

$$\rho = \frac{s^2 - n}{n(n-1)}$$

Based on equation 2.5 we can already define a lower bound for the average correlation ρ. This is because we know $s^2 > 0$. Starting from the total variance expression:

$$s^2 = 2\sum_{i<j} \sigma_i \sigma_j \rho + \sum_{i=1}^{n} \sigma_i^2 > 0$$

$$2\sum_{i<j} \sigma_i \sigma_j \rho > -\sum_{i=1}^{n} \sigma_i^2$$

$$\rho > -\frac{\sum_{i=1}^{n} \sigma_i^2}{2\sum_{i<j} \sigma_i \sigma_j}$$

Since $\rho \leq 1$ must also hold, we obtain the final bounds:

$$-\frac{\sum_{i=1}^{n} \sigma_i^2}{2\sum_{i<j} \sigma_i \sigma_j} \leq \rho \leq 1$$

Corollary 2. *An important observation can be made from expression 2.4 by applying the triangle inequality. Specifically, a matrix T exists if and only if:*

$$max\{\sigma_{i_{max}} - \sum_{i \neq i_{max}} \sigma_i, 0\} \leq s \leq \sum_{i=1}^{n} \sigma_i \tag{2.6}$$

Where σ_{max} denotes the largest value of σ_i. Note that if $\sigma_i = 1$ for all i, we have that

$$max\{2 - n, 0\} \leq s \leq n$$

Hence, the matrix T is constructed iteratively by constructing row vectors t_i such that their weighted sum eventually satisfies the condition in equation 2.4. In order to do this we must determine feasible lengths that the cumulative sums $\sum_{i=1}^{j} t_i$ can attain, ensuring that the final sum has norm s.
The problem thus reduces to generating a matrix T with unit-norm rows $t_i \in \mathbb{R}^n$, such that the weighted sum $\sum_{i=1}^{n} \sigma_i t_i$ has squared norm equal to a prescribed value s^2 (see equation 2.4).
Directly sampling such vectors is difficult due to the global constraint on their weighted sum. Which is why we instead build vectors t_i iteratively, ensuring at each step that the partial sums remain consistent with the eventual goal of reaching norm s. Recall from table 2.1 that $l_i := \left\| \sum_{j=1}^{i} \sigma_j t_j \right\|$ for $i = 1, \ldots, n$
These lengths l_i must be carefully constructed so that it is geometrically possible to reach final length s. If you have two vectors, the length of their sum must lie between the difference and the sum of their individual lengths. That is, the new length l_i must be somewhere between $|l_{i-1} - \sigma_i|$ and $l_{i-1} + \sigma_i$.

The following theorem characterizes what constraints must be applied to such a sequence of vectors in order to satisfy the final constraint on their lengths.

Theorem 2. *Suppose that $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_n$. Let $1 \leq k \leq n-2$ be an integer, and let $l_1, \ldots, l_k$ be non-negative real numbers satisfying:*

$$l_1 = \sigma_1 \quad \textit{and} \quad |l_i - l_{i-1}| \leq \sigma_i \leq l_i + l_{i-1} \quad \textit{for all } 2 \leq i \leq k.$$

Then the following two statements are equivalent:

1. *There exist non-negative real numbers $l_{k+1},\dots,l_n$ with $l_n = s$ such that*
$$|l_i - l_{i-1}| \le \sigma_i \le l_i + l_{i-1} \quad \textit{for all } 2 \le i \le n.$$
2.
$$|s - l_k| \le \sum_{i=k+1}^{n} \sigma_i \quad \textit{and} \quad s + l_k \ge \sigma_n - \sum_{i=k+1}^{n-1} \sigma_i.$$

Proof. This proof is reproduced from Tuitman et al. [14, Proof of Theorem 2].

$(1 \Rightarrow 2)$ If $l_{k+1},...,l_n$ satisfying the conditions exist, then

$$|s - l_k| = |l_n - l_k| \le |l_{k+1} - l_k| + \cdots + |l_n - l_{n-1}| \le \sigma_{k+1} + \cdots + \sigma_n,$$
$$s + l_k = (l_n + l_{n-1} + (l_k - l_{k+1}) + \cdots + (l_{n-2} - l_{n-1}) \ge \sigma_n - (\sigma_{k+1} + \cdots + \sigma_{n-1}).$$

$(2 \Leftarrow 1)$ We construct the l_i inductively for $k+1 \le i \le n-1$, by choosing

$$l_i \in \left[\max|l_{i-1} - \sigma_i|, s - \sum_{j=i+1}^{n} \sigma_j, \sigma_n - \sum_{j=i+1}^{n-1} \sigma_j - s, \min l_{i-1} + \sigma_i, s + \sum_{j=i+1}^{n} \sigma_j \right]$$

□

Corollary 3. *If $l_1,...,l_{i-1}$ have been constructed such that*
$$|l_j - l_{j-1}| \le \sigma_j \le l_j + l_{j-1} \textit{ for all } 2 \le j < i$$
Then the next l_i must have length within the interval :
$$l_i \in \left[max\left\{ |l_{i-1} - \sigma_i|, s - \sum_{j=i+1}^{n} \sigma_j, \sigma_n - \sum_{j=i+1}^{n-1} \sigma_j - s \right\}, min\left\{ l_{i-1} + \sigma_i, s + \sum_{j=i+1}^{n} \sigma_j \right\} \right] \tag{2.7}$$
If $\sigma_j = 1$ for all $j = 1,...,n$ the interval will be
$$l_i \in [max\{|l_{i-1} - 1|, s - n + i + 1, 3 - n + i - s\}, min\{l_{i-1} + 1, s + n - i - 1\}] \tag{2.8}$$

In order to make sure that the new cumulative sum has length l_i we must carefully control how t_i aligns with u_{i-1}. Inner product $\langle t_i, u_{k-1} \rangle$ determines how much of t_i is pointing in the direction of u_{i-1}. If t_i points towards u_{i-1} the resulting sum is longer, if it points away the sum is shorter.

The following lemma gives a useful condition: In order for the new cumulative vector $u_i = u_{i-1} + \sigma_i t_i$ to have desired length l_i, the inner product between t_i and u_{i-1} must be equal to a specific value. This value is determined by l_i, l_{i-1} and σ_i.

Lemma 2. *For $u_i = \sum_{j=1}^{i} \sigma_j t_j$ we have*
$$\|u_k\| = \ell_k \iff \langle t_k, u_{k-1} \rangle = \frac{\ell_k^2 - \ell_{k-1}^2 - \sigma_k^2}{2\sigma_k}$$

Proof. This proof is reproduced from Tuitman et al. [14, Proof of lemma 2]. We have

$$\begin{aligned} \| u_k \|^2 &= \| u_{k-1} + \sigma_k t_k \|^2 \\ &= \| u_{k-1} \|^2 + \| \sigma_k t_k \|^2 + 2\langle \sigma_k t_k, u_{k-1} \rangle \\ &= l_{k-1}^2 + \sigma_k^2 + 2\sigma_k \langle t_k, u_{k-1} \rangle \end{aligned}$$

□

Clearly the length of the sum of t_i must eventually be s, however the construction of t_i must be so that the length of the vectors gradually attains s and not immediately, in addition the vectors created cannot be collinear.

To achieve this, t_i is written as the sum of two components:
$$t_i = z_i + y_i \tag{2.9}$$
where:

- z is a component in the direction of the previous sum u_{i-1} and contributes to the length
- y is a random component orthogonal to u_{i-1}

Constructing directional component z_i
The vector z is constructed such that adding t_i to u_{i-1} results in a new vector $u_i = u_{i-1} + t_i$ with squared norm equal to a given value l_i^2.
In order for this to be achieved a scalar α must be found such that $z_i = \alpha u_{i-1}$ increases the length of u_{i-1} to the target length l_i.
By lemma 2 we know the following:

$$\| u_i \| = l_i \iff \langle t_i, u_{i-1} \rangle = \frac{\ell_i^2 - \ell_{i-1}^2 - \sigma_i^2}{2\sigma_i}$$

Hence we can use this result to find α

$$\begin{aligned} \langle t_k, u_{i-1} \rangle &= \langle \alpha u_{i-1} + y, u_{i-1} \rangle \\ &= \alpha \| u_{i-1} \|^2 \\ &= \frac{l_i^2 - l_{i-1}^2 - \sigma_i^2}{2\sigma_i} \\ &\Rightarrow \alpha = \frac{l_i^2 - l_{i-1}^2 - \sigma_i^2}{2\sigma_i \| u_{i-1} \|^2} \end{aligned}$$

Therefore:

$$z_i = u_{i-1} \cdot \left(\frac{l_i^2 - l_{i-1}^2 - \sigma_i^2}{2\sigma_i \cdot \| u_{i-1} \|^2} \right) \tag{2.10}$$

If $\sigma_i = 1$, for all $i = 1, ..., n$ then

$$z_i = u_{i-1} \cdot \left(\frac{l_i^2 - l_{i-1}^2 - 1}{2 \cdot \| u_{i-1} \|^2} \right) \tag{2.11}$$

Constructing orthogonal component y_i
y_i is a random vector orthogonal to u_{i-1}, To construct the orthogonal component y_i , we begin by generating a random vector $x \in \mathbb{R}^n$. Then removing its projection onto u_{i-1} the resulting vector is orthogonal to the current direction:

$$y_i = x - \frac{\langle x, u_i \rangle}{\| u_i \|^2} \cdot u_1$$

This guarantees that $y_i \perp u_{i-1}$. To ensure that the full vector $t_i = z_i + y_i$ has unit length, y_i must be rescaled, hence it is multiplied by

$$\frac{\sqrt{1 - \| z_i \|^2}}{\| y_i \|}$$

This vector is necessary to ensure that the new direction of t_i is not collinear with the previous vector. If this were the case then the matrix T would have linearly dependent row, resulting in a low-rank matrix and a correlation matrix $C = TT'$ that is not positive-semi-definite. Furthermore, strong co-linearity between vectors result in large pairwise inner product, hence satisfying a target average correlation constraint becomes difficult.

Computing t_i in this manner guarantees that vectors remain on the unit sphere while contributing the exact amount required per iteration to control the total variance. Once the row vectors t_i have been constructed, the correlation matrix $C = TT'$ is obtained and has the desired average correlation ρ.

2.2.3. Implementation and results

Algorithm 2 Generation of series of admissible lengths

Given: $\sigma_1, \ldots, \sigma_n$ and s such that $0 \leq s \leq n$
Generate: $l_1, \ldots, l_n$
$l_1 \leftarrow 1$
for $i := 2$ to $n-1$ **do**
    $l_i \leftarrow$ random in $[\max\{s - \sum_{j=i+1}^{n-1} \sigma_j, \sigma_n - \sum_{j=i+1}^{n-1} \sigma_j - s, |l_{i-1} - \sigma_i|\}, \min\{s + \sum_{j=i+1}^{n} \sigma_j, l_{i-1} + \sigma_i\}]$
end for
$l_n \leftarrow s$
Return: $l_1, \ldots, l_n$

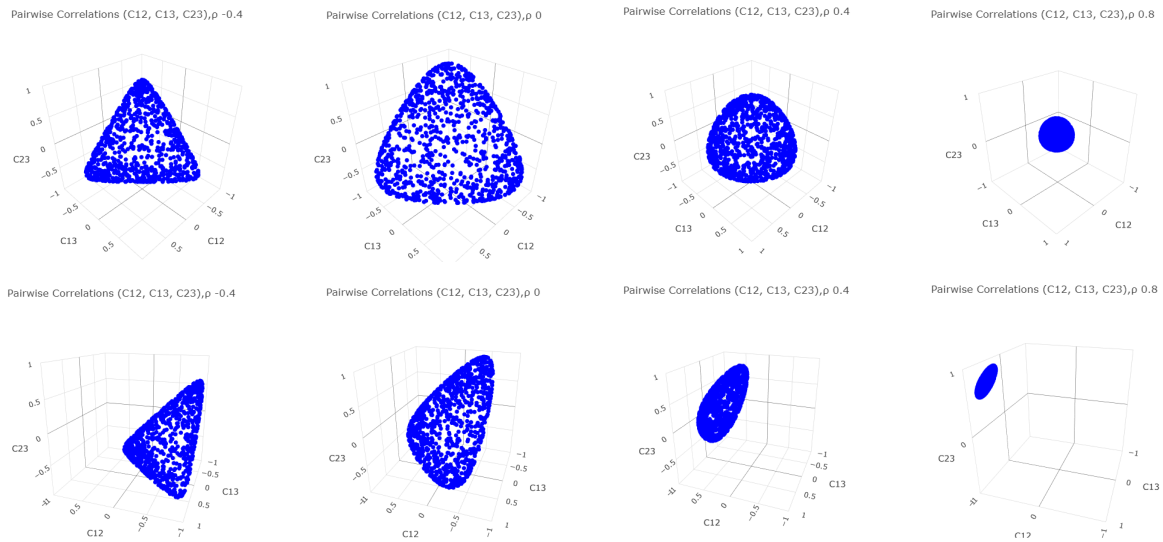
Algorithm 3 Generation of row vectors t_i

Generate: $t_1, \ldots, t_n$, given $l_1, \ldots, l_n$
$t_1 \leftarrow$ random in $\mathbb{R}^n$ of length 1
$u_1 \leftarrow t_1$
for $i := 2$ to n **do**
    $x \leftarrow$ random in $\mathbb{R}^n$ of length 1
    $y \leftarrow x - u_{i-1} \cdot \dfrac{\langle x, u_{i-1} \rangle}{\|u_{i-1}\|^2}$
    $z \leftarrow u_{i-1} \cdot \dfrac{l_i^2 - l_{i-1}^2 - \sigma_i^2}{2\sigma_i \|u_{i-1}\|^2}$
    $t_i \leftarrow z + \left(\dfrac{\sqrt{1 - \|z\|^2}}{\|y\|} \right) y$
    $u_i \leftarrow u_{i-1} + \sigma_i t_i$
end for
Return: $t_1, \ldots, t_n$

The computation times for generating correlation matrices using this implementation are summarized in Table 2.3 below. Note that the first two rows, are only available for $n = 5$ and $n = 10$, since the lower bound for the feasible average correlation is $\rho > -\frac{1}{n-1}$. It can be seen that the computation time increases with matrix dimension n. This is expected, as the number of off-diagonal elements grows quadratically (specifically $\frac{n(n-1)}{2}$), and the algorithm must ensure that the resulting matrix satisfies the average correlation and the positive semi-definite constraint. The target average correlation ρ does not appear to affect the computational time.

Table 2.3: Computation time (ms) for generating a correlation matrix using the algorithm, for selected matrix sizes and average correlation values ρ.

ρ	**Computation times for different matrix dimensions n**										
	5	**10**	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
−0.25	0.0353	–	–	–	–	–	–	–	–	–	–
−0.10	0.0360	0.0697	–	–	–	–	–	–	–	–	–
0.00	0.0341	0.0709	0.1969	0.2878	0.3911	0.5441	0.6866	0.9363	1.1308	1.4663	1.6152
0.20	0.0567	0.0835	0.1626	0.2737	0.3925	0.5597	0.7153	0.8739	1.2491	1.4702	1.5730
0.40	0.0347	0.0732	0.1633	0.3493	0.3932	0.5885	0.6985	0.8791	1.2223	1.4800	1.6479
0.60	0.0483	0.0775	0.1611	0.2606	0.4136	0.5925	0.7218	0.9688	1.1502	1.3776	1.7249
0.80	0.0343	0.0716	0.1693	0.2800	0.4420	0.5449	0.7058	0.9024	1.1302	2.3528	1.6491



$\rho = -0.4$, corresponding to the red surface in figure 1.1d

$\rho = 0.0$ corresponding to the green surface in figure 1.1d

$\rho = 0.4$, corresponding to the orange surface in figure 1.1d

$\rho = 0.8$, corresponding to the brown surface in figure 1.1d

Figure 2.3: 3D scatter plots of pairwise correlations (C_{12}, C_{13}, C_{23}) sampled from running the algorithm 1000 times $\rho \in \{-0.4, 0.0, 0.4, 0.8\}$. Each column shows a view from above and a rotated view of the same slice.

Figure 2.3 shows 1000 3×3 correlation matrices generated for different average correlations. The scatter plots illustrate the structure of the feasible space. We indeed find that all of the matrices lie on a slice of the elliptical tetrahedron as described for Figure 1.1.

From the scatter plots, it is clear that as ρ increases, the feasible region for the matrices becomes smaller. For high average correlation the method requires vectors to align more closely, which restricts the freedom when sampling vectors and can make it harder to maintain independence among the the rows of T. Furthermore, high average correlations corresponds to a smaller slice of the elliptical tetrahedron. As a result, this alignment can lead to violation of positive semi-definiteness or the shrinking of the feasible region can cause numerical instability.

2.2.4. Example

In this section a concrete example of generating a 3x3 matrix with average correlation $\rho = 0.4$ is presented

Step 1: Compute the target norm s. The overall norm s of the vector sum is derived from equation 2.2:

$$s^2 = 2 \cdot \rho \cdot \frac{n(n-1)}{2} + n = 2 \cdot 0.4 \cdot \frac{3 \cdot 2}{2} + 3 = 5.4, \quad \text{so} \quad s = 2.324$$

Step 2: Compute the intermediate norms l_i. We begin with $l_1 = \| t_1 \| = 1$. The value of l_2 must be chosen to satisfy the constraints imposed by the constraints from Theorem 2. It is sampled from the interval 2.8:

$$l_2 \in [\max\{0, s-3+2+1, 3-3+2-s\}, \min\{1+1, s+3-2+1\}]$$

$$l_2 \in [1.324, 2]$$

In this example, $l_2 = 1.988$ is chosen uniformly from this interval, we set $l_3 = s = 2.324$ by definition. We now aim to find a set of unit vectors t_1, t_2, t_3 such that:

$$u_i = \sum_{j=1}^{i} t_j, \quad \text{with} \quad \|u_i\| = l_i, \quad \text{and} \quad \|t_i\| = 1.$$

Step 3: Construct t_1. We begin by sampling a random unit vector from $\mathbb{R}^n$ to begin the algorithm:

$$t_1 = \begin{bmatrix} 0.962 \\ 0.2196 \\ 0.161 \end{bmatrix}$$

So $u_1 = t_1$, and by construction $\|u_1\| = l_1 = 1$.

Step 4: Construct t_2. As we saw in the previous section, each t_i is constructed as $t_i = z + y$. Hence to construct t_2, we must find the orthogonal component y and deterministic component z. To find y we generate another random unit vector from $\mathbb{R}^n$:

$$x_2 = \begin{bmatrix} 0.947 \\ -0.054 \\ -0.317 \end{bmatrix}$$

Then this vector is projected onto u_1 to remove its component in the direction of u_1, ensuring orthogonality. The projection coefficient is:

$$p = \frac{x_2 \cdot u_1}{\|u_1\|^2} = 0.848$$

$$y = x_2 - p \cdot u_1 = \begin{bmatrix} 0.131 \\ -0.240 \\ -0.453 \end{bmatrix}$$

This vector y is then normalized:

$$y_{\text{norm}} = 0.408 \cdot y = \begin{bmatrix} 0.0535 \\ -0.098 \\ -0.185 \end{bmatrix}$$

Next, we compute the scalar α and vector z with equation 2.11:

$$\alpha = \frac{l_2^2 - l_1^2 - \sigma_2^2}{2 \cdot \sigma_2 \cdot \|u_1\|^2} = \frac{1.988^2 - 1^2 - 1^2}{2 \cdot 1 \cdot 1^2} = \frac{3.952 - 2}{2} = 0.976$$

$$z = \alpha \cdot u_1 = 0.976 \cdot \begin{bmatrix} 0.962 \\ 0.2196 \\ 0.161 \end{bmatrix} = \begin{bmatrix} 0.939 \\ 0.214 \\ 0.157 \end{bmatrix}$$

$$t_2 = z + y_{\text{norm}} = \begin{bmatrix} 0.939 \\ 0.214 \\ 0.157 \end{bmatrix} + \begin{bmatrix} 0.0535 \\ -0.098 \\ -0.185 \end{bmatrix} = \begin{bmatrix} 0.993 \\ 0.117 \\ -0.028 \end{bmatrix}$$

Step 5: Construct t_3. The third vector is constructed similarly, ensuring orthogonality with the previous vector sum $u_2 = t_1 + t_2$, and ensuring that $\|u_3\| = s$. A random unit vector x_3 is sampled, projected orthogonally to u_2, and combined with a scaled u_2 component (via z_3) to form t_3. The process mirrors that used for t_2.

$$x_3 = \begin{bmatrix} 0.184 \\ 0.193 \\ 0.008 \end{bmatrix}, y_3 = \begin{bmatrix} -0.0893 \\ 0.146 \\ 0.945 \end{bmatrix}, z_3 = \begin{bmatrix} 0.111 \\ 0.019 \\ 0.008 \end{bmatrix}, t_3 = \begin{bmatrix} 0.018 \\ 0.193 \\ 0.964 \end{bmatrix}$$

$$T = \begin{bmatrix} 0.962 & 0.2196 & 0.161 \\ 0.993 & 0.1166 & -0.0279 \\ 0.0182 & 0.170 & 0.985 \end{bmatrix}$$

This concludes the construction. The final matrix is then given by stacking the t_i vectors as rows of t_i and computing $C = TT^T$. This guarantees a symmetric, positive semidefinite matrix with 1's on the diagonal and average off-diagonal correlation ρ.

$$C = \begin{bmatrix} 1 & 0.9765 & 0.213 \\ 0.9765 & 1 & 0.0104 \\ 0.213 & 0.0104 & 1 \end{bmatrix}$$

Indeed when checking the average correlation value we do have $\rho = 0.4$.

2.2.5. Properties of the matrices generated with the algorithm by Tuitman et al

This subsection examines the behaviour of the correlation matrices C generated by the algorithm described in Section 2. The primary focus lies on the case $n = 3$, although higher-dimensional cases are also considered. In particular, we study how the individual off-diagonal entries C_{ij} vary with different values of the average correlation ρ. Additionally, we assess their variance and dependence structure.

The analysis proceeds as follows:

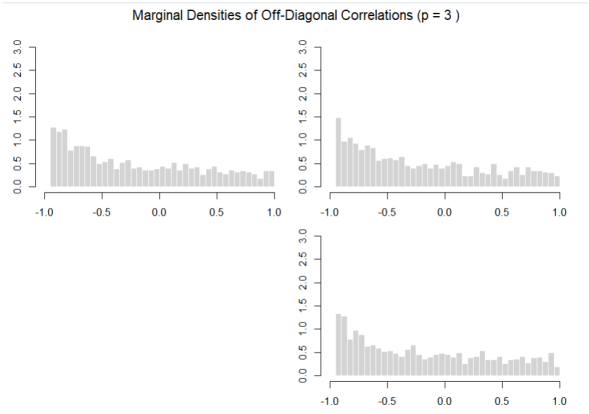
1. We simulate 1000 random correlation matrices using the Algorithms 2 and 3 with average $\rho \in \{-0.2, 0.0, 0.4, 0.8\}$.
2. From each simulation, we extract and store the off-diagonal elements C_{ij}.
3. The marginal densities, box plots of the entries and summary statistics including means, variances, and pairwise correlations are computed and presented in Table 2.5.

We expect that for higher values of ρ, the marginal distributions for each C_{ij} become increasingly restricted. Furthermore, in higher dimensions, the feasible region for the correlation matrix becomes more confined, leading to stronger structural constraints on the entries of C.

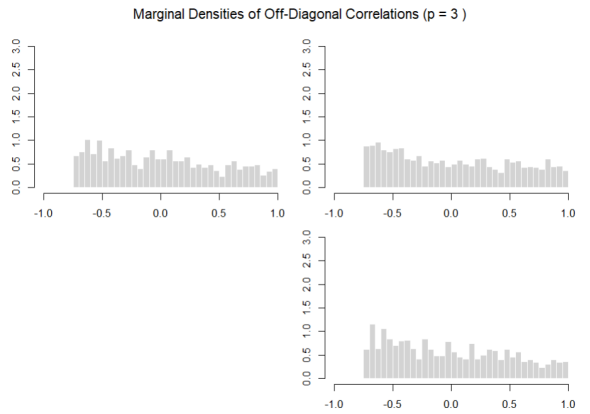
Figure 2.4 shows the marginal distributions of the off-diagonal elements C_{12}, C_{13} and C_{23} for correlation matrices generated using the algorithm. Several trends are evident from these plots:

- For small values of ρ such as -0.2 and 0.0, Figures 2.4a and 2.4b, the marginal distributions appear wider. This indicates high geometric flexibility, a wide variety of C_{ij} values can satisfy the global average correlation constraint. This agrees with the intuition presented for Figure 2.3, where the feasible region of valid correlation matrices is larger for lower values of ρ.
- For larger values of ρ, see Figure 2.4c for $\rho = 0.8$, the marginal densities become sharply peaked and concentrated near the upper bound of the correlation interval. This reflects the algorithm's reduced flexibility: as ρ increases, there are fewer admissible configurations that satisfy all constraints.

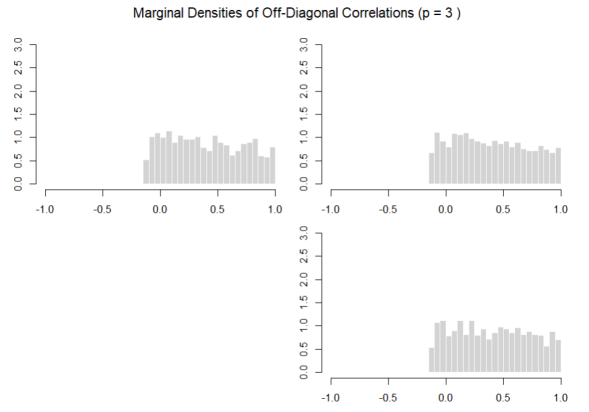
The behaviour described above is consistent with what was expected, as ρ increases, the space of admissible matrices narrows and hence the diversity of correlation values reduces. This observation is further supported by the box plots shown in Figure 2.5, where the range and variance of the correlation values visibly shrink as ρ increases.



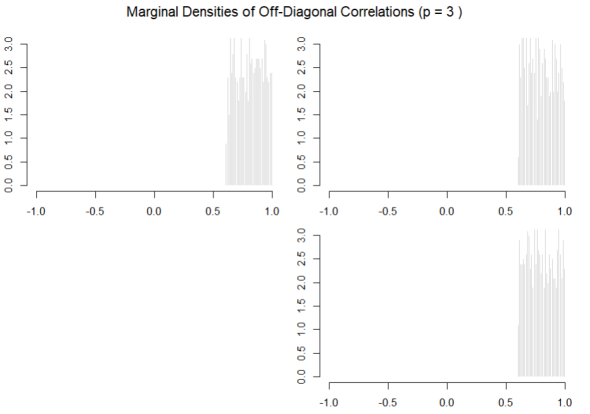
(a) Marginal densities for $\rho = -0.2$



(b) Marginal densities for $\rho = 0.0$



(c) Marginal densities for $\rho = 0.4$



(d) Marginal densities for $\rho = 0.8$

Figure 2.4: Marginal densities at each position C_{ij} for $C \in \mathbb{M}_{3\times 3}(\mathbb{R})$ under different values of average correlation $\rho \in \{-0.2, 0.0, 0.4, 0.8\}$.

The marginal densities do not appear to follow a specific distribution, this can be attributed to the geometry and constraints applied by this method. The matrices are generated by building T such that the constraint $s^2 = \sum_{i,j} C_{ij}\sigma_i\sigma_j$ is satisfied. Early vectors are sampled with more freedom while later vectors are tightly constrained to achieve the final norm s. As a result the matrix elements are sampled under different conditions, this construction introduces asymmetry, or $n > 3$. We can nevertheless test whether these matrices resemble a uniform or shifted beta distribution. In three dimensions we can test only on C_{12}, given that the elements are invariant under permutations. This is because the constraints that form this region the matrices lie on are:

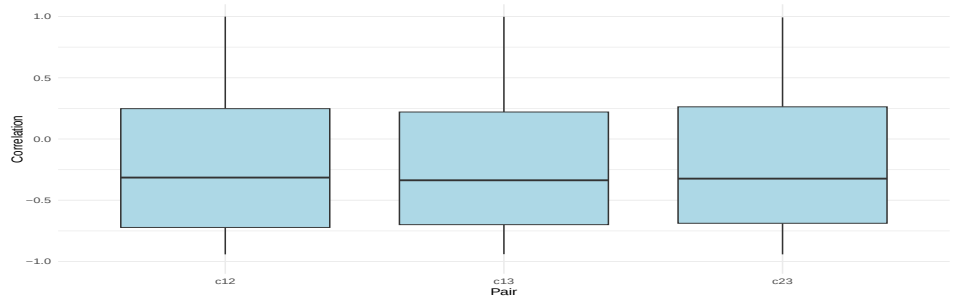
$$\det(C) = 1 + 2C_{12}C_{13}C_{23} - C_{12}^2 - C_{13}^2 - C_{23}^2 \geq 0$$

$$C_{12} + C_{13} + C_{23} = 3\rho.$$

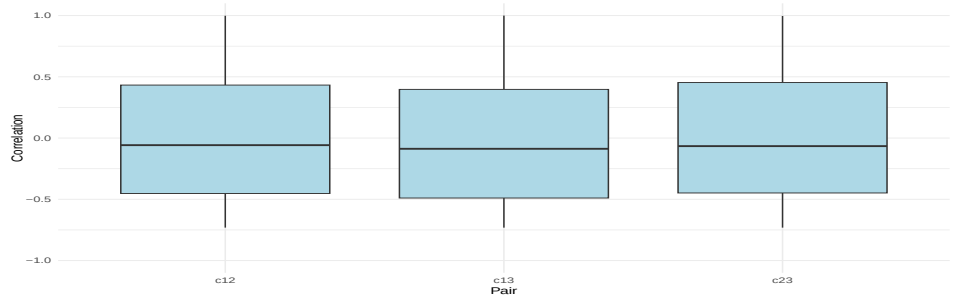
These constraints are both symmetric in the variables C_{12}, C_{13} and C_{23} which means they are invariant under permutations, so the distribution of C_{ij} will be identical over the plane. The Kolmogorov-Smirnov test was applied for the goodness-of-fit test to the uniform distribution on the empirical support, and the best-fit beta distribution via maximum likelihood estimation. The results are presented in Table 2.4. Here we see that uniform and beta distributions are rejected across all μ values. While the distributions in Figure 2.4 seem to be fairly uniformly spread the KS test strongly rejects uniformity. For $\mu = -0.2$ the beta fit is better as it captures the left skew distribution seen in 2.4 better than uniform. However for high μ the beta p-value gets larger, in Figure 2.4 we also no longer see a skewed behaviour but rather stronger peak near 1. Here the uniform fits slightly better but still is rejected. Hence from these tests we do not see a distribution that fits these margins exactly, which reflects the underlying geometric constraints of the construction.

μ	Beta Fit (α, β)	KS p-value (Beta)	KS p-value (Uniform)
-0.2	$(0.89,\ 1.30)$	0.0016	$< 2.2 \times 10^{-16}$
0.0	$(1.51,\ 1.38)$	5.3×10^{-5}	1.6×10^{-10}
0.4	$(4.21,\ 1.79)$	7.6×10^{-8}	3.0×10^{-5}
0.8	$(16.26,\ 1.75)$	3.0×10^{-8}	2.9×10^{-3}

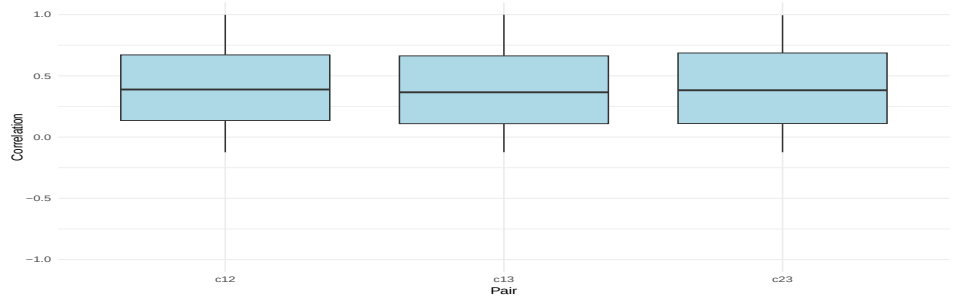
Table 2.4: Goodness-of-fit testing of marginal distributions of C_{12} for different values of μ.



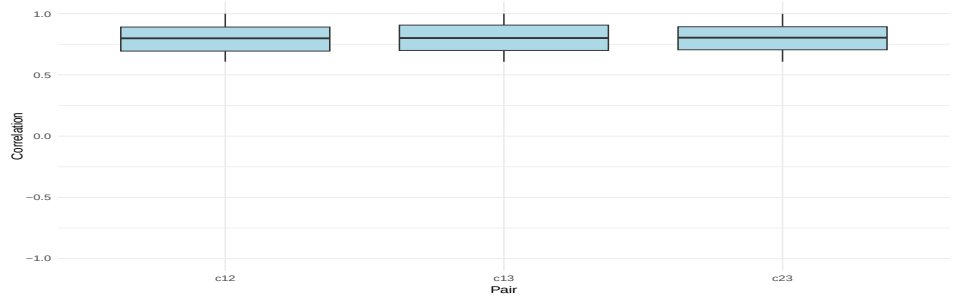
(a) Box plots for $\rho = -0.2$



(b) Box plots for $\rho = 0.0$



(c) Box plots for $\rho = 0.4$



(d) Box plots for $\rho = 0.8$

Figure 2.5: Box plots of the entries c_{ij} for correlation matrices $C \in \mathbb{M}_{3\times 3}(\mathbb{R})$ generated under different target average correlations ρ.

Table 2.5 summarizes the statistical properties of the 1000 simulations for various average correlations. For each ρ, we report the empirical averages and variances of off-diagonal correlation elements, and the correlation between these elements. This further gives us an insight on the structure of the generated matrices, and how this varies for different values of ρ.

Table 2.5: Simulation summary per average correlation ρ

Avg ρ	**Expected Correlation**			**Variance**			**Correlation between Elements**		
	c_{12}	c_{13}	c_{23}	c_{12}	c_{13}	c_{23}	$C_{12,13}$	$C_{12,23}$	$C_{13,23}$
-0.2	-0.202	-0.187	-0.211	0.314	0.330	0.314	-0.499	-0.491	-0.510
0	0.014	-0.018	0.003	0.244	0.258	0.264	-0.487	-0.499	-0.514
0.4	0.400	0.409	0.391	0.106	0.107	0.102	-0.487	-0.499	-0.514
0.8	0.797	0.805	0.798	0.013	0.013	0.013	-0.500	-0.482	-0.518

- The first column shows the expected value of C_{ij} . It is clear that the algorithm successfully generates matrices with given average correlation.
- The variance of each off diagonal elements reflects what we also saw in figures 2.4 and 2.5. For large ρ the variance is very low, for $\rho = 0.8$ even as low as 0.013. At $\rho = -0.2$ the variance is much higher, 0.31, which is what we saw in the plots.
- Another notable observation is that the correlation between off-diagonals is approximately the same per value of ρ. In addition they are negative, this reflects that if one correlation increases, the other two must decrease in order to satisfy the average correlation constraint. As mentioned earlier, C_{12}, C_{13} and C_{23} are invariant under permutations, as result the off-diagonal elements are statistically equivalent. We will show below why the correlation between entries of the correlation matrices in case $n = 3$ is -0.5.

  Let C_{12}, C_{23} and C_{13} be random variables. We have

$$\text{Var}(C_{12} + C_{23} + C_{13}) = \text{Var}(C_{12}) + \text{Var}(C_{13}) + \text{Var}(C_{23}) + 2\text{Cov}(C_{12}, C_{23}) + 2\text{Cov}(C_{12}, C_{13}) + 2\text{Cov}(C_{13}, C_{23}) \tag{2.12}$$

  As mentioned above, each off-diagonal element has the same variance and covariance due to symmetry. Denote as r the correlation between these random variables and τ^2 their variances.

$$\text{Corr}(C_{ij}, C_{kl}) = r, \quad \text{Var}(C_{ij}) = \tau^2, \quad \text{Cov}(C_{ij}, C_{kl}) = r\tau^2 \quad \text{for all } i, j \neq (k, l).$$

  Then we get

$$\text{Var}(C_{12} + C_{23} + C_{13}) = 3\tau^2 + 6r\tau^2.$$

  In addition we know that the sum of the off-diagonal elements is always equal to 3ρ by definition of the algorithm, so the variance of the sum of these elements is equal to zero. Hence

$$3\tau^2 + 6r\tau^2 = 0.$$

  Solving this gives :

$$r = -\frac{1}{2}.$$

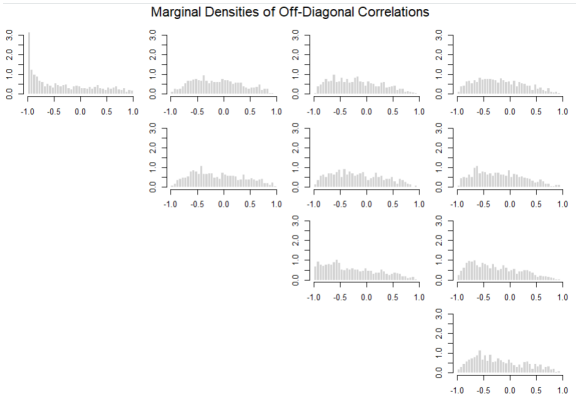
Unlike for $n =$, in higher dimensions, there is no general closed form expression for the correlation between off-diagonal elements of the matrix $C = TT'$. As the number of variables increases, the dependence structure among entries becomes more complex due to the high dimensional geometry and global norm constraint imposed. In Table 2.6 some correlations between off-diagonal elements across 1000 simulations are presented for $n = 5$. While elements that share an index often display positive empirical correlation, due to shared dependence on a common unit vector, but this does depend on how they are aligned. Disjoint entries tend to exhibit negative correlation due to the fixed average constraint. This pattern arises because a change in a shared vector (for example t_3) can simultaneously effect elements C_{34} and C_{35}, as we see in the table these have positive correlations. Whereas disjoint entries are not directly coupled but remain indirectly constrained by the fixed average correlation. In order to achieve the target average correlation an increase in one entry often requires a decrease in others. However these are empirical tendencies rather than strict rules.

For example, in Table 2.6 the correlation between C_{12} and C_{13} is -0.154, illustrating that shared indices do not always imply positive correlation.

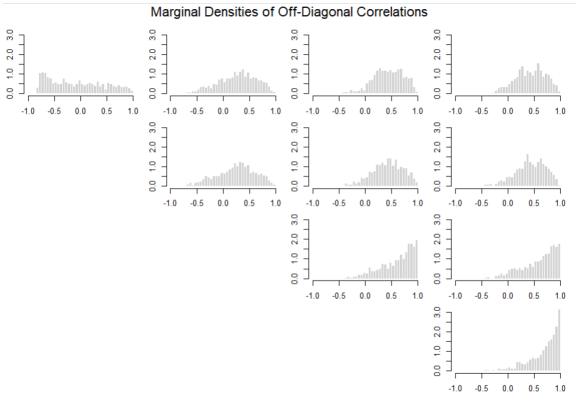
C_{ij}	C_{kl}	Correlation
C_{12}	C_{13}	-0.154
C_{12}	C_{34}	-0.681
C_{12}	C_{35}	-0.707
C_{34}	C_{35}	0.399
C_{35}	C_{45}	0.395

Table 2.6: Selected correlations between off-diagonal entries of a 5×5 correlation matrix, illustrating both weak and strong dependencies.

Figure 2.6 shows the marginal densities for $n = 5$, when $\rho = -0.2$ and 0.4. When $\rho = -0.2$ the marginal distributions seem relatively symmetric and broadly spread, consistent with the geometric interpretation discussed above. For $\rho = 0.4$ the marginal distributions are more skewed and increasingly asymmetrical, this indicated indeed that the feasible region is much more constrained for large values of ρ. Only a small section of the feasible region satisfies the high average correlation constraint, and also adheres to the positive semi-definite constraint.



(a) Marginal densities for $C \in \mathbb{M}_{5\times5}(\mathbb{R})$, $\rho = -0.2$



(b) Marginal densities for $C \in \mathbb{M}_{5\times5}(\mathbb{R})$, $\rho = 0.4$

Figure 2.6: Marginal densities at each position c_{ij} for $C \in \mathbb{M}_{5\times5}(\mathbb{R})$ for different values of ρ.

In higher dimensions, the feasible region we are sampling from is a hyperplane of dimension $\frac{n(n-1)}{2} - 1$. The resulting intersection is a complex space which is difficult to visualize. However, we can investigate some properties this space must have. While the average constraint treats all off-diagonal elements equally, the positive semi-definite condition does not for $d > 3$. The behaviour seen in Table 2.6 for correlations between off-diagonal elements can also be explained as a result of the positive semidefinite constraint. The condition requires that all determinants of sub-matrices are non-negative. The result of this is that certain elements C_{ij} appear more frequently, which means these are more tightly constrained. This is often the case for C_{ij} that share an index. Disjoint pairs appear in larger sub-matrices and are hence less directly linked.

The deeper global structure can also be evaluated, when ρ is higher this means the vectors t_i become more aligned, resulting in similar directions. While this increases average correlation, this also means that eigenvalues become smaller as they become more collinear. This results in the fact that these matrices lie closer to the boundary of the feasible region as they are close to breaking the positive semi-definite constraint. In addition, for high-dimensional convex bodies, such as the hyperplane the matrices lie on, it is known that most of the volume concentrates near the boundary. [1] As a result, variables introduced earlier in the construction have more freedom, while those assigned later are more constrained and tend to lie closer to the edges of the feasible region. This often corresponds to matrices where elements with higher indices have with more extreme entries.This behaviour can be viewed empirically in Figure 2.6, where elements involving higher indices show stronger skewness. This suggests that the marginal distributions concentrate near the extremes not only due to the positive semi-definite and average constraint, but also as a consequence of the

geometric properties of the high dimensional hyperplane. For smaller or negative average correlations, this geometric effect is less pronounced. In these cases the intersection of the average-correlation hyperplane and the space of correlation matrices defines a larger feasible region. So although the volume concentrates near the boundary, the larger feasible region means there is also volume in the interior. This suggests an additional explanation, beyond the effect of the positive semi-definite constraint, for why marginal distributions with lower ρ are less skewed than those for higher ρ.

2.2.6. Comparison of the results for matrices generated with and without average constraint by SRD parametrization

The SRD parametrization generates correlation matrices using random unit vectors. This construction guarantees positive semi-definiteness, but there is no further structure imposed on the distribution of the matrices. As a result we see in Figure 2.1 that the marginal distributions are uniformly distributed and are uncorrelated, reflecting the independence of the sampled vectors. In contrast the approach by Tuitman et al. [14] modifies the parametrization to impose an average correlation. This is achieved by iteratively constructing vectors t_i such that the norm of their weighted sum is fixed, indirectly enforces the desired average correlation via a variance constraint. The matrices generated under this constraint have very different structural behaviour, the matrices are located on a slice of the feasible region. The feasible region becomes highly constrained for high values of ρ and n. The contrast in behaviour of the matrices highlights an important trade-off. The unconstrained method allows for a fast and unbiased method to sample correlation matrices, but lacks control over the dependency structure of the matrix. The Tuitman et al. [14] method allows for control over the average correlation of the matrices, however this comes at the cost of geometric complexity. This motivated the exploration of alternative methods that can balance structural control with flexibility. In the next chapter, a different approach based on partial correlation c-vine parametrisation is explored.

3

Generating correlation matrices using partial correlation parametrization

An alternative method to construct valid correlation matrices is to parametrize them in terms of partial correlations, using C-vine decomposition. In this approach, the full correlation matrix is uniquely determined by a structured sequence of partial correlations, each conditioned on a growing set of variables. This parametrization translates to a recursive construction: by sampling the partial correlations from appropriately chosen distributions on $(-1,1)$, and combining them via a specific recursion, the resulting matrix is a valid correlation matrix. A key advantage is that the difficult positive definite constraint is replaced by a sequence of simpler, local constraints that are straightforward to satisfy in simulation.This chapter introduces the method by Lewandowski et al. [11], the algorithm and illustrative examples. We then examine an extension proposed by Joe and Kurowicka [10], which shows how the first moment of the off-diagonal elements can be controlled by sampling partial correlations from appropriately chosen Beta distributions. The properties and implications of these extensions are also discussed.

Notation:

- C_{ij}: the correlation coefficient between random variables X_i and X_j.
- $C_{ij;S}$: the correlation between X_i and X_j fixing variables indexed in $S \subset \{1,\dots,n\} \setminus \{i,j\}$.
- If $S = \emptyset$ then $C_{ij;\emptyset} = C_{ij}$

The set of partial correlations used in this method can be described by the partial correlation C-vine. A vine on n variables is a nested set of connected trees $T_1,...,T_{n-1}$ where the edges of tree T_i are the nodes of tree $T_{i+1}, i = 1,...,n-2$[11].

Definition 2: [16] Partial correlation $C_{ij;1,...,k}$ is defined as the Pearson (linear) correlation between residuals from the linear regressions

$$X_i = a_i + b_i'(X_1,...,X_k) + \varepsilon_i, \qquad X_j = a_j + b_j'(X_i,...,X_k) + \varepsilon_j$$

that is,

$$C_{ij;1,...,k} = \mathrm{corr}(\varepsilon_i, \varepsilon_j).$$

The partial correlation $\rho_{i,j;1...k}$ measures the linear association between X_i and X_j after removing the linear effects of $X_1,...,X_k$ from both variables. This partial correlation between X_i and X_j with X_k, can be calculated as follows:

$$C_{ij;k} = \frac{C_{ij} - C_{ik}C_{jk}}{\sqrt{(1-C_{ik}^2)(1-C_{jk}^2)}} \tag{3.1}$$

The partial correlations between X_i and X_j, with X_L fixed where $L \subset \{1,...,n\}$ and $k,i,j \notin L$ can be calculated recursively:

$$C_{ij;kL} = \frac{C_{ij;L} - C_{ik;L}C_{jk;L}}{\sqrt{(1-C^2_{ik;L})(1-C^2_{jk;L})}}. \tag{3.2}$$

The goal is to construct a valid $n \times n$ correlation matrix $C = (C_{ij})_{1 \le i,j \le n}$. This will be achieved by parametrizing this matrix using a partial correlation C-vine with the following independent parameters:

$$\begin{array}{cccll} C_{12} & C_{13} & \cdots & C_{1n} & \text{tree } 1 \\ & C_{23;1} & \cdots & C_{2n;1} & \text{tree } 2 \\ & & \ddots & \vdots & \vdots \\ & & & C_{n-1,n;1,\ldots,n-2} & \text{tree } n-1 \end{array}$$

There is a one-to-one relationship between the set of partial correlations defined along the C-vine and the entries of a valid correlation matrix [2]. Joe [9] introduced a recursive formula, which allows partial correlations to be transformed into standard correlations in a structured manner. The algorithm for generating these correlation matrices is based on these recursive relationships, which define how to construct a full correlation matrix from a structured set of partial correlations equations (see Section 2 of Lewandowski et al. [11] for detailed derivations).

The following recursion is built using equation 3.2 , see Joe and Kurowicka [10] section 2.1. The partial correlations of row l given a set $S = 1,...,n$ can be computed as follows, with $j > l$ for $n > l$:

$$C_{lj;1\cdots l-2} = C_{lj;1\cdots l-1}\sqrt{1-C^2_{l-1,l;1\cdots l-1}}\sqrt{1-C^2_{l-1,j;1\cdots l-1}} + C_{l-1.l;1\cdots l-2}C_{l-1,j;1\cdots l-2}$$

For $1 \le k < l-2$,

$$C_{lj;1\cdots k} = C_{lj;1\cdots k+1}\sqrt{1-C^2_{k+1,l;1\cdots l}}\sqrt{1-C^2_{k+1,l;1\cdots l}} + C_{k+1,l;1\cdots k}C_{k+1,j;1\cdots k}$$

and

$$C_{lj} = C_{lk;1}\sqrt{1-C^2_{1l}}\sqrt{1-C^2_{1j}} + C_{1l}C_{1j}$$

Hence putting this all together we get:

$$C_{\ell j} = \left\{ \sum_{i=1}^{\ell-1} C_{i\ell;1,\ldots,i-1}\, C_{ij;1,\ldots,i-1} \prod_{k=1}^{i-1} \sqrt{1-C^2_{k\ell;1,\ldots,k-1}}\sqrt{1-C^2_{kj;1,\ldots,k-1}} + C_{\ell j;1,\ldots,\ell-1} \prod_{k=1}^{\ell-1} \sqrt{1-C^2_{k\ell;1,\ldots,k-1}}\sqrt{1-C^2_{kj;1,\ldots,k-1}} \right\}$$

An advantage of this approach is that the partial correlations can be sampled independently from known distributions, such as a Beta distribution transformed to the interval $(-1,1)$. In principle any distribution for partial correlations can be used but it is very convenient to pick a Beta distribution for this purpose. If parameters of these beta distributions are chosen carefully then it was shown in Joe and Kurowicka [10] that the resulting joint density of correlations in the correlation matrix is as follows:

$$f_{C_n}(C) = \prod_{\ell=1}^{n-1} \prod_{j=\ell+1}^{n} f_{C_{\ell j;1:\ell-1}}(C_{\ell j;1:\ell-1}) \cdot \prod_{\ell=1}^{n-1} \prod_{j=\ell+1}^{n} (1-C^2_{\ell j;1:\ell-1})^{-(n-\ell-1)/2} \tag{3.3}$$

If the partial correlations follow a symmetric beta distribution on (-1,1) with the correct parameters we achieve the LKJ-distribution for elements of C. This specific beta distribution $Z = 2W - 1$ has support on $(-1,1)$, with $W \sim \text{Beta}(0,1)$ and parameter $\alpha_k = \alpha + \frac{n-k-1}{2}$, where k represents a tree in a C-vine. Then the resulting correlation matrix follows the LKJ-distribution where the density is proportional to

$$\det(C)^{\alpha-1} \tag{3.4}$$

Furthermore, the marginal distributions follow a known distribution, namely symmetric Beta on (-1,1) with parameter $\alpha - 1 + \frac{n}{2}$. Hence, in the case where $\alpha = 1$, which by 3.4 corresponds to a uniform distribution over the space of correlation matrices, has marginal distributions distributed as Beta$(\frac{n}{2}, \frac{n}{2})$. This allows for straightforward computation of quantities such as the expectation and variance of the matrix entries, which is not always the case if we sample partial correlations from other distributions.

Expectation of matrix elements

If $W \sim \text{Beta}(a,b)$ on $(0,1)$, then the transformation $Z = 2W-1$ maps the support to $(-1,1)$ and the expectation and variance are:

$$\mathbb{E}(C_{ij}) = \frac{2a}{a+b} - 1$$
$$\text{Var}(C_{ij}) = \frac{4ab}{(a+b)^2(a+b+1)} \tag{3.5}$$

So if indeed have that the correlation matrices are uniformly distributed, $\alpha = 1$, then $a = b = \frac{n}{2}$ and we find:

$$\mathbb{E}(C_{ij}) = 0, \text{Var}(C_{ij}) = \frac{1}{n+1}.$$

However, in the case that the matrices are not uniformly distributed but still LKJ, the margins are $\text{Beta}(\alpha - 1 + \frac{n}{2}, \alpha - 1 + \frac{n}{2})$ which gives:

$$\mathbb{E}(C_{ij}) = 0, \quad \text{Var}(C_{ij}) = \frac{1}{2\alpha + n - 1}. \tag{3.6}$$

If the distributions of partial correlations are chosen to have different distributions than the case of LKJ distribution then the marginal distributions are not known. However, we can still compute the expectations of C_{ij} recursively.

Since partial correlations in C-vine parametrization are independent and we assume that partial correlations in the same tree have the same distribution, the following observation can be made. In tree 1 all correlations are independent. Due to the recursive formula there are always only two variables from one tree level used in the computation of give correlation. W denote them as $X_i = C_{i\ell;1\ldots i-1}$ and $Y_i = C_{ij;1\ldots i-1}$, where ℓ is the tree level and $j > \ell$ is suppressed.
Then,

$$C_{\ell j} = Y_\ell \prod_{k=1}^{\ell-1} \sqrt{1-X_k^2}\sqrt{1-Y_k^2} + \sum_{i=1}^{\ell-1} X_i Y_i \prod_{k=1}^{i-1} \sqrt{1-X_k^2}\sqrt{1-Y_k^2}$$

Let $\mathbb{E}(X_i) = \mu_i$, $\mathbb{E}(X_i^2) = \nu_i$ $\mathbb{E}((1-X_i)^{\frac{1}{2}}) = \gamma_i$, $\mathbb{E}(X_i(1-X_i^2)^{\frac{1}{2}}) = \eta_i$, and similarly for Y_i.
Then the expectation of $C_{\ell j}$ for $j > \ell$ is:

$$\mathbb{E}(C_{\ell j}) = \mu_\ell \prod_{k=1}^{\ell-1} \gamma_k^2 + \sum_{i=1}^{\ell-1} \mu_i^2 \prod_{k=1}^{i-1} \gamma_k^2$$

This can also be written as: (starting with $\mathbb{E}(C_{12}) = \mu_1$)

$$\mathbb{E}(C_{\ell j}) = \mathbb{E}(C_{\ell-1,j}) + ([\mu_{\ell-1}^2 - \mu_{\ell-1}] + \mu_\ell \gamma_{\ell-1}^2) \prod_{i=1}^{\ell-2} \gamma_i^2, \ell \geq 2, \tag{3.7}$$

The second moment and variances of the margins can also be computed, but will not further be discussed in this paper. To see this computation see section 3.3 in the paper by Joe and Kurowicka [10].

Overview of the method

1. Sample partial correlations $C_{ij;1,\ldots,i-1}$ for $1 \leq i < j \leq n$ from a distribution of choice. For the LKJ-distribution over matrices, choose symmetric Beta with parameter $\alpha_l = \alpha + \frac{n-k-1}{2}$.
2. Use the recursive formula to compute C_{ij} from the partial correlations.
3. Assemble the matrix $C = (C_{ij})$

Example of generating a 3×3 Correlation Matrix

We begin by constructing a 3×3 correlation matrix using the partial correlation C-vine method, we will use beta distributions for the partial correlations. In this case, we require two levels of partial correlations:

$$\begin{array}{ll} C_{12} & C_{13} \quad \text{tree 1} \\ & C_{23;1} \quad \text{tree 2} \end{array}$$

To transform $C_{23;1}$ into C_{23} we can rewrite equation 3.1 to get:

$$C_{23} = C_{23;1}\sqrt{1-C_{12}^2}\sqrt{1-C_{13}^2} + C_{12}C_{13}$$

Step 1: Determine Beta parameters per level

Here we will be using $\alpha = 1$

Tree k	$a = b = \alpha + \frac{n-k-1}{2}$
1	1.5
2	1

Step 2: Determine the partial correlations

The following partial correlations were sampled from the corresponding Beta distributions:

C_{12}	C_{13}	$C_{23;1}$
0.786	-0.166	0.676

Step 3: Computing correlation matrix entries

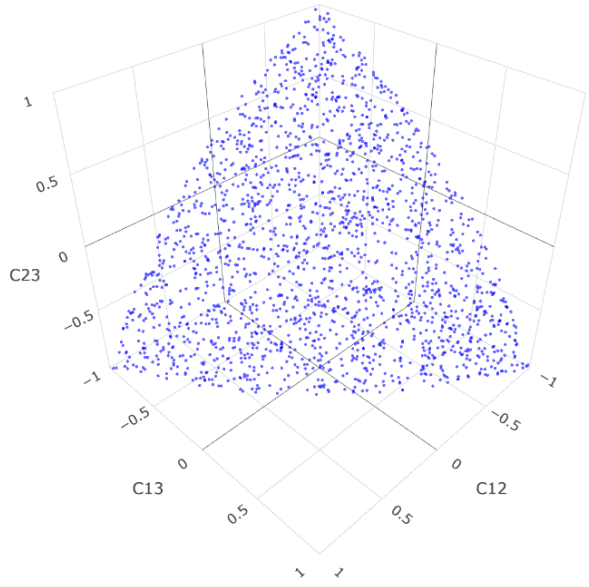
Using the recursive formulas for the C-vine structure, we compute the remaining off-diagonal entry:

$$\begin{aligned} C_{23} &= C_{23;1}\sqrt{(1-C_{12}^2)(1-C_{13}^2)} + C_{12}C_{13} \\ &= 0.676 \cdot \sqrt{(1-0.786^2)(1-(-0.166)^2)} + 0.786 \cdot (-0.166) \\ &= 0.676 \cdot \sqrt{0.383 \cdot 0.972} - 0.131 \\ &= 0.676 \cdot 0.610 - 0.131 = 0.282 \end{aligned}$$

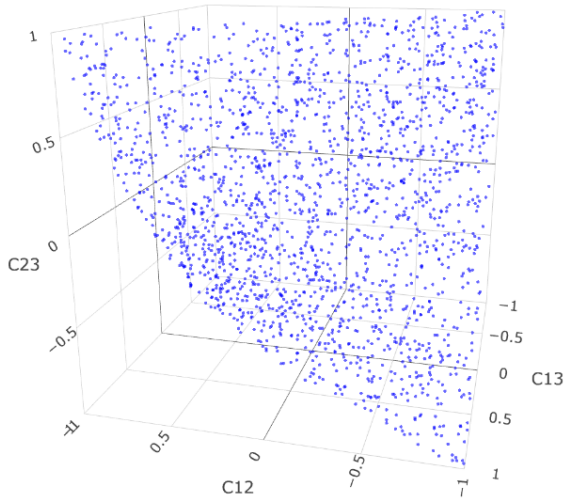
Final 3×3 correlation matrix

$$C = \begin{pmatrix} 1 & 0.786 & -0.166 \\ 0.786 & 1 & 0.282 \\ -0.166 & 0.282 & 1 \end{pmatrix}$$

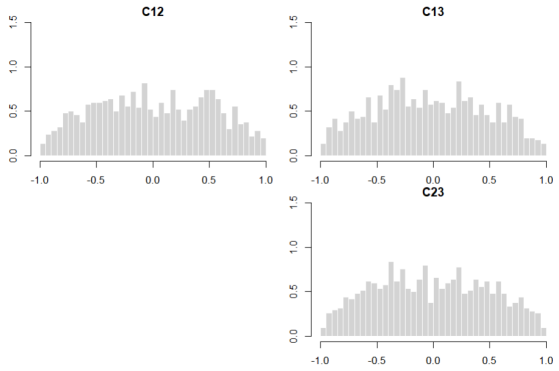
Figure 3.1 shows 1000 randomly generated 3×3 correlation matrices with $\alpha = 1$, sampled from the LKJ distribution using the partial correlation C-vine method. Compared to Figure 1.1 from the introduction, it is evident that the sampled points are uniformly distributed within the feasible region. This confirms that,when $\alpha = 1$ and the Beta distributions are chosen according to the LKJ-parametrization, the partial correlation approach yields samples that are uniformly distributed over the space of valid correlation matrices. Furthermore, the marginal distributions illustrate that the off-diagonal entries are indeed distributed as Beta($\frac{n}{2}, \frac{n}{2}$), consistent with the properties of the LKJ distribution.



(a) 1000 randomly generated correlation matrices



(b) 1000 randomly generated correlation matrices



(c) Beta($\frac{3}{2}$, $\frac{3}{2}$) distributed marginals

Figure 3.1: 1000 correlation matrices generated by the LKJ-method, $\alpha = 1$

Extending to a 4×4 Correlation Matrix

To demonstrate how this nesting works we can extend the 3x3 example we did to four variables, we introduce additional partial correlations at tree level 1, 2, and 3:

C_{12}	C_{13}	C_{14}	tree 1
	$C_{23;1}$	$C_{24;1}$	tree 2
		$C_{34;12}$	tree 3

Step 1: Determine Beta parameters per level and partial correlations

We continue using the same values from example 3,and we sample C_{14},$C_{24;1}$ and $C_{34;12}$ from the beta distribution with parameters shown in the table below. Again, $\alpha = 1$.

Tree k	$a = b = \alpha + \frac{n-k-1}{2}$
1	2.0
2	1.5
3	1.0

The complete set of sampled partial correlations is:

C_{12}	C_{13}	C_{14}	$C_{23;1}$	$C_{24;1}$	$C_{34;12}$
0.786	-0.166	-0.312	0.676	0.084	0.468

Step 2: Computing correlation matrix entries

We now compute the remaining off-diagonal entries using the recursive formulas:

Compute C_{24}:

$$\begin{aligned} C_{24} &= C_{24;1}\sqrt{(1-C_{12}^2)(1-C_{14}^2)} + C_{12}C_{14} \\ &= 0.084 \cdot \sqrt{(1-0.786^2)(1-(-0.312)^2)} + 0.786 \cdot (-0.312) \\ &= 0.084 \cdot \sqrt{0.382 \cdot 0.903} - 0.245 \\ &= 0.084 \cdot 0.587 - 0.245 = 0.049 - 0.245 = -0.196 \end{aligned}$$

Compute C_{34}:

$$\begin{aligned} C_{34} &= C_{34;12}\sqrt{(1-C_{23;1}^2)(1-C_{24;1}^2)(1-C_{13}^2)(1-C_{14}^2)} \\ &\quad + C_{23;1}C_{24;1}\sqrt{(1-C_{13}^2)(1-C_{14}^2)} + C_{13}C_{14} \\ &= 0.468 \cdot 0.688 + 0.0568 \cdot \sqrt{0.878} + 0.052 \\ &= 0.322 + 0.053 + 0.052 = 0.427 \end{aligned}$$

Final 4×4 correlation matrix

$$C = \begin{pmatrix} 1 & 0.786 & -0.166 & -0.312 \\ 0.786 & 1 & 0.282 & -0.196 \\ -0.166 & 0.282 & 1 & 0.427 \\ -0.312 & -0.196 & 0.427 & 1 \end{pmatrix}$$

3.1. Properties of the C-vine Partial correlation parametrization

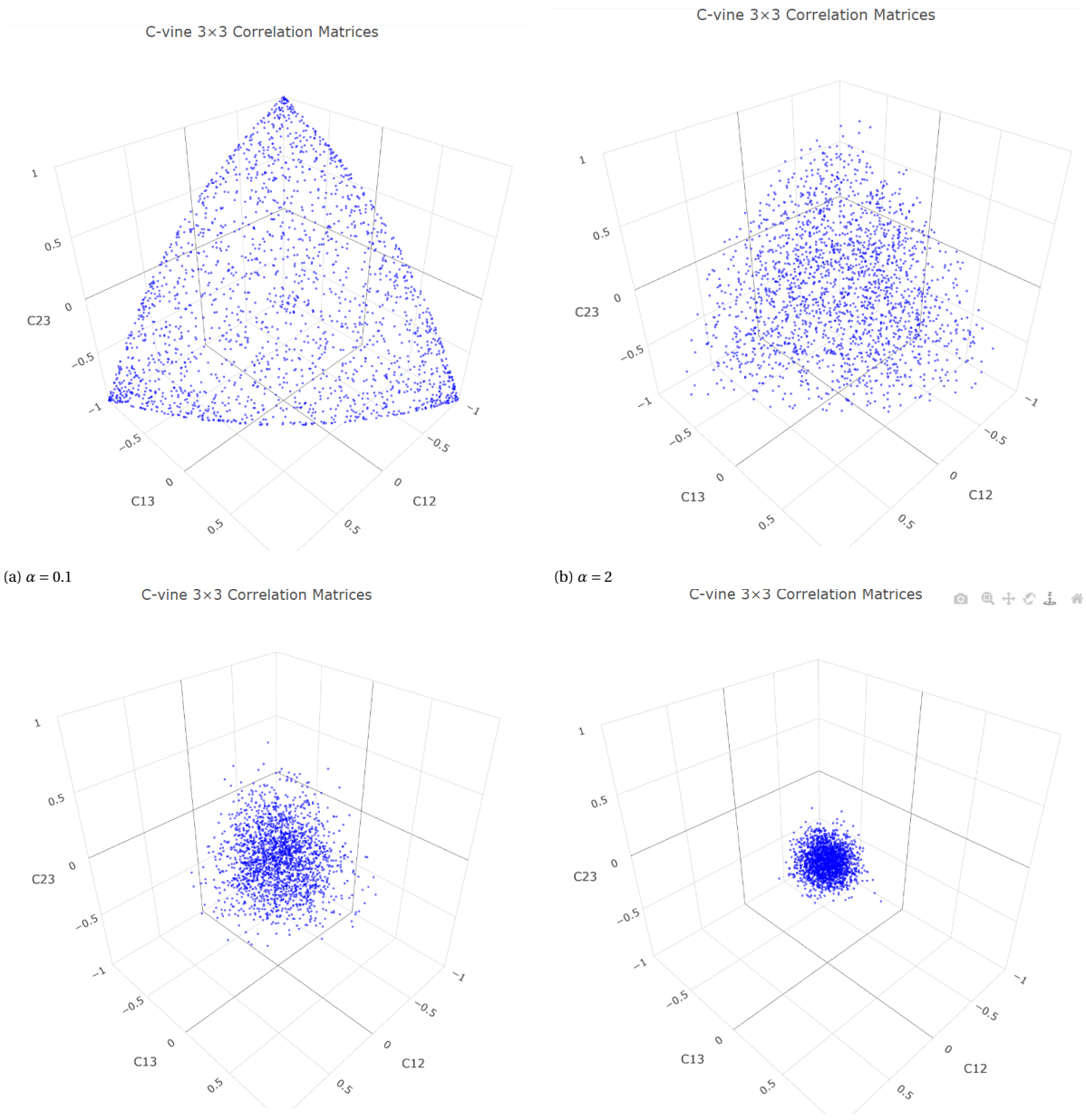
In this section, we explore the statistical properties of the correlation matrices generated by this method. We begin by examining the expectation of the off-diagonal elements. Following this we explore how the choice of the LKJ- parameter α influences the distribution of samples over the space of valid correlation matrices. We then consider the case outside LKJ distribution, hence when we sample partial correlations from non-symmetric Beta distributions.

3.1.1. Scatter plots and marginal distributions of sampled correlation matrices with LKJ distribution

To illustrate how the LKJ parameter α influences the structure of sampled correlation matrices, we visualize the distribution of the off-diagonal elements in three-dimensions. We sample 1000 matrices from the C-vine parametrization method using Symmetric Beta distributions such that the matrices have LKJ distribution.

- In Figure 3.2a , when $\alpha = 0.1$, the samples are concentrated near the boundaries of the elliptical tetrahedron. This reflects that fact that matrices with strong positive or negative correlations are favoured, hence clustering near the edges of the feasible region.
- When $\alpha = 2$, shown in Figure 3.2b a slight shift is seen, while the matrices are still spread out, they begin to cluster more toward the origin. Hence these matrices tend to have slightly weaker correlations.
- This pattern continues in the following plots, clearly in Figure 3.2c, where $\alpha = 10$, the matrices are concentrated in the middle and the distribution strongly favours weak correlations. Finally, in Figure 3.2d where $\alpha = 50$ the matrices are so concentrated in the middle that most of the generated matrices are close to the identity matrix.

The parameter α serves to control the correlation strength. Low $\alpha < 1$ encourages strong dependence and for $\alpha > 1$ the matrices cluster around the identity matrix.



(a) $\alpha = 0.1$

(b) $\alpha = 2$

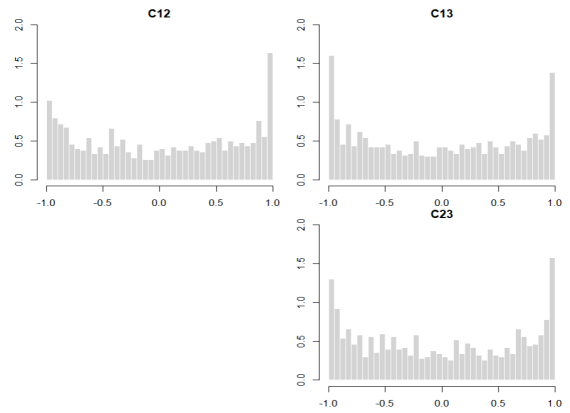
(c) $\alpha = 10$

(d) $\alpha = 50$

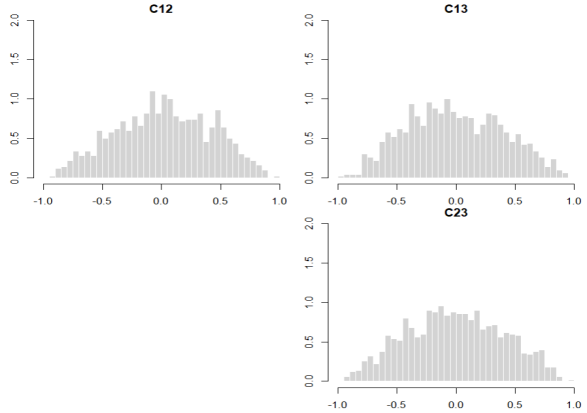
Figure 3.2: 3D scatter plots of pairwise correlations (C_{12}, C_{13}, C_{23}) sampled from running the LKJ-algorithm, 1000 times. Each subplot shows samples generated with a different Beta parameter α.

As with the previous method, we can again examine the marginal distributions of the correlations generated. The marginal densities agree with the patterns observed in Figures 3.2.

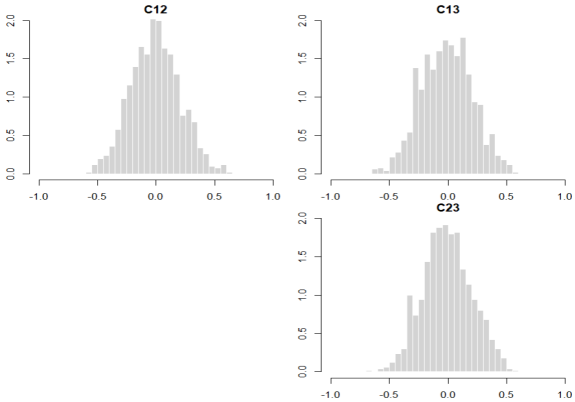
- When $\alpha < 1$, Figure 3.3a, the distributions exhibit increased mass near the extremes, as expected for such beta distributions, reflecting a higher likelihood of strong positive or negative correlations. This behaviour is consistent with Figure 3.3a, where samples clustered around the boundary of the elliptical tetrahedron.
- For larger values of α, Figures 3.3c and 3.3d, the densities concentrate around zero, indicating that the matrices are close to the identity matrix. Additionally, from equation 3.5 that the variance decreases as α increases, which can also be seen in Figures 3.2 and 3.3, which can be expected looking at the variance equation in 3.6.



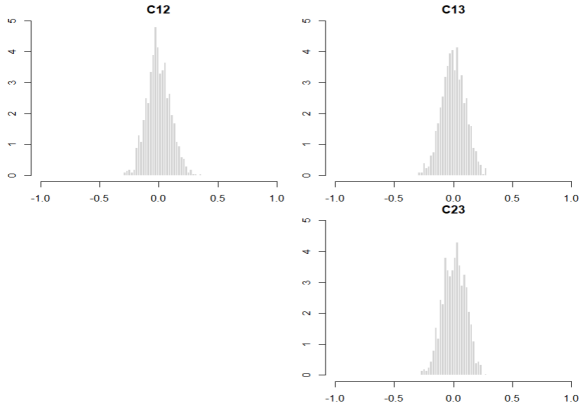
(a) Marginal distributions for $\alpha = 0.1$



(b) Marginal distributions for $\alpha = 2$



(c) Marginal distributions for $\alpha = 10$



(d) Marginal distributions for $\alpha = 50$

Figure 3.3: Marginal distributions at each position C_{ij} for 1000 matrices $C \in \mathbb{M}_{3\times 3}(\mathbb{R})$ for different α.

Next we show an example of the method which does not conform to the LKJ distribution.

3.1.2. Example using asymmetric Beta partial correlations

Let us simulate partial correlations on C-vine which are asymmetric Beta distributions. The following correlations and partial correlations are sampled independently from Beta$(2,5)$. We get:

$$C_{12} = -0.412$$
$$C_{13} = -0.538$$
$$C_{23;1} = 0.105$$

Using the same recursive formula as Example 3 we find $C_{23} = 0.302$. Hence the resulting correlation matrix is:

$$C = \begin{bmatrix} 1 & -0.412 & -0.538 \\ -0.412 & 1 & 0.302 \\ -0.538 & 0.302 & 1 \end{bmatrix}.$$

The resulting matrix C is symmetric, positive definite, and lies within the feasible region of valid correlation matrices, however this correlation matrix does not follow the LKJ-distribution. The recursion guarantees validity, but the distribution over the space of correlation matrices depends on the distributions used for the partial correlations. Furthermore the marginals are no longer identically. This can be seen in the marginal distributions in Figure 3.4. This can be seen in Table 3.1, this shows the empirical quantiles at 5%, 25%, 75% and 95%, so that we can compare the spread per C_{ij}. While the difference is subtle, it is clear that since C_{12} and C_{13} are directly derived from the sampled partials, they exhibit symmetric marginal behaviour. However, C_{23} is computed via the recursive formula, as a result the distribution is different. Particularly in the 5th and 95th percentiles we see a difference. Hence the asymmetry in the marginal distributions

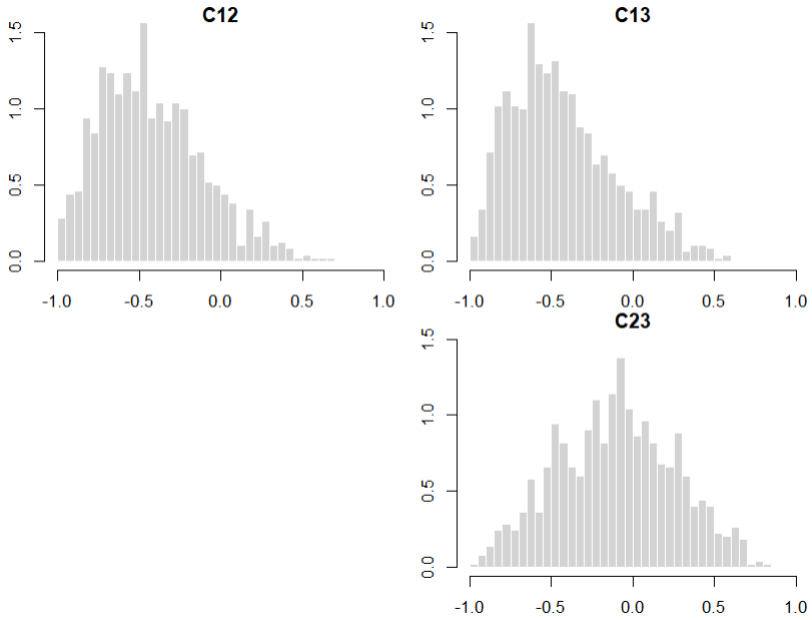


Figure 3.4: Marginal distribution of off-diagonal elements for correlation matrices sampled by partial correlations from Beta(2,5)

and the variation between them show that sampling from non-symmetric beta distributions breaks the exchangeability and symmetry observed when using LKJ-distribution parameters.

Quantile	C_{12}	C_{13}	C_{23}
5%	−0.748	−0.761	−0.515
25%	−0.375	−0.344	−0.201
50%	−0.020	−0.007	0.000
75%	0.333	0.325	0.255
95%	0.712	0.698	0.524

Table 3.1: Empirical quantiles of marginal distributions for C_{12}, C_{13}, and C_{23} from sampled correlation matrices.

The method of sampling partial correlations and using recursive formula still works in this case but it is very difficult to derive information about the margins beyond the level 1 correlations. For example, C_{23} is distributed as follows, derived from equation 3: [10]:

$$f_{C_{23}}(c_{23}) = \int_{-1}^{1}\int_{-1}^{1} (1-c_{12}^2)^{-\frac{1}{2}} (1-c_{13}^2)^{-\frac{1}{2}} f_{C_{23;1}}\Big(\frac{c_{23}-c_{12}c_{13}}{\sqrt{1-c_{12}^2}\sqrt{1-c_{13}^2}}\Big) f_{C_{12}}(c_{12})\, f_{C_{13}}(c_{13})\, dc_{12}\, dc_{13}. \tag{3.8}$$

However this analytical approach to finding $f_{C_{23}}$ is intractable in closed form. Instead, statistical properties of the margins are best obtained through recursive formulas, such as the expectation formula 3.7 and expressions for the second moment and variances as detailed in section 3.3 of Joe and Kurowicka [10].

3.1.3. Permutation-based symmetrization of correlation matrix entries

Above we saw that the algorithm causes complicated marginal distributions for the correlations generated using the recursive formula with partials sampled without LKJ parameters. If one wants that all correlations in the matrix have the same marginal distribution then the rows and columns of matrix C can be permuted. In that case the distribution of correlations in the correlation matrices in three dimensions 3.3 becomes:

$$f_C(c) = f_{C_{12}}(c_{12}) f_{C_{13}}(c_{13}) f_{C_{23;1}}(c_{23;1})(1-c_{12}^2)^{\frac{1}{2}}(1-c_{13})^{\frac{1}{2}}$$

and the marginal density of each correlation is:

$$f_C(c) = \frac{\left[f_{C_{12}}(c) + f_{C_{13}}(c) + f_{C_{23}}(c)\right]}{3}.$$

If we now again calculate the percentiles for the same example as Table 3.1 (matrices generated using Beta(2,5) partials), we find that indeed the elements C_{ij} are now distributed the much more similarly.

Quantile	C_{12}	C_{13}	C_{23}
5%	−0.687	−0.686	−0.689
25%	−0.291	−0.323	−0.314
50%	−0.004	−0.011	−0.015
75%	0.289	0.281	0.281
95%	0.682	0.678	0.650

Table 3.2: Empirical quantiles of marginal distributions for permuted C_{12}, C_{13}, and C_{23} with partial correlations sampled from Beta(2,5)

For these matrices the expectation is: Let C^π be the correlation matrix after a random permutation π , then the expectation for any off-diagonal element C^π_{ij} is:

$$\mathbb{E}\left[C^\pi_{ij}\right] = \frac{2}{n(n-1)} \sum_{\ell=1}^{n-1} (n-\ell)\,\mathbb{E}\left[C_{\ell,\ell+1}\right] \tag{3.9}$$

3.2. Fixing the expectation of each correlation

In Chapter 2, we investigated a method that allows to generate matrices with fixed average correlation. In this section, we extend our analysis of the method based on partial correlations by examining how we can control the expectation of marginal distributions of correlations in correlation matrix, which also allows to control the expectation of average correlation. We start with fixing the expectation of each off-diagonal element of the correlation matrix. Then the properties of matrices generated with this approach are investigated.

Using recursion an expression can be found that holds if all expectations of correlations in the matrix are equal (to μ_1)[10]. We observe that the following relationship is satisfied between expectations of partial correlations in different levels in this case.

$$\mu_\ell = \mu_{\ell-1}\left[\frac{1-\mu_{\ell-1}}{\gamma^2_{\ell-1}}\right], \ell \in 3,4,....$$

Naturally the following condition has to be satisfied: $-1 < \mu_{\ell-1}\left[\frac{1-\mu_{\ell-1}}{\gamma^2_{\ell-1}}\right] < 1$.

So if we have that $C_{i\ell;1\ldots i-1} \sim Beta(a, b_\ell)$ then $\mu_{a,b_\ell} = \frac{a-b_\ell}{a+b_\ell}$ and $\gamma_{a,b_\ell} = \mathbb{E}[(1-X^2_\ell)^{\frac{1}{2}}]$. When $b > a$ then $\mu_{a,b} < 0$ and we have the following proposition from Joe and Kurowicka [10].

Proposition 1. Let $X \sim$ Beta(a,b) on $(-1,1)$ with $a > 0, b > 0$. For all $a \le b$,

$$\frac{(1-\mu_{a,b})}{\gamma^2_{a,b}} \ge 1,$$

where $\mu_{a,b} = \frac{b-a}{a+b}$ and

$$\gamma_{a,b} = \mathbb{E}[(1-X)^{1/2}] = \frac{2B(a+0.5, b+0.5)}{B(a,b)}.$$

What this reveals is that if the initial partial correlations are sampled from a Beta distribution with $a \le b$ (i.e., with negative mean), the factor $\frac{1-\mu_{a,b}}{\gamma^2_{a,b}}$ is always at least 1, and typically larger. As the recursion proceeds, each successive expectation μ_ℓ can become more negative in magnitude. For sufficiently large ℓ, it is possible for the recursion to reach or exceed the lower boundary, i.e., $\mu_{a,b_\ell} \le -1$, which is not admissible.

In other words, in these cases it is impossible to find parameters for the partial correlations such that all entries in the resulting correlation matrix have the same expectation, when that target expectation is sufficiently negative. So, the recursive structure imposes a geometric constraint on the range of achievable expectations as the dimension increases.

To ensure that the expected partial correlations remain consistent across levels of the C-vine, the algorithm adaptively constructs a sequence of Beta distributions with parameters (a, b_k) for each tree level k. The partial correlations are drawn from a transformed Beta distribution on $(-1,1)$ via

$$W_k \sim 2Z - 1, \quad Z_k \sim \text{Beta}(a, b_k),$$

which yields an expected value

$$\mu_k := \mathbb{E}[Z_k] = \frac{2a}{a+b_k} - 1.$$

We want to ensure that this expected value remains approximately constant across tree levels, i.e.,

$$\mu_1 \approx \mu_2 \approx \cdots \approx \mu_k,$$

In particular, when setting two consecutive means equal to each other we get the following equation:

$$\mu_k^2 - \mu_k + \mu_{k+1} \cdot \gamma_k^2 \approx 0,$$

where $\gamma_k = \mathbb{E}[(1-W_k)^{1/2}]$ can be computed using properties of the Beta distribution.

To find a value of b_{k+1} that satisfies this equation we hence need to minimize the following function:

$$g(b) := \left(\mu_k^2 - \mu_k + \left(\frac{2a}{a+b} - 1\right) \cdot \gamma_k^2\right)^2,$$

The algorithm minimizes this numerically. This is accomplished using R's `nlm` function:

```
outsolve = nlm(giter, bvec[i])
bvec[i+1] = outsolve$estimate
```

Here, `giter` implements the function $g(b)$ defined above, and `nlm` finds the value of b that minimizes it, starting from an initial guess b_k.

The result is a sequence $b_1, b_2, \ldots, b_k$ such that each Beta distribution $\text{Beta}(a, b_k)$ produces partial correlations that ensure the expectation of the resulting random correlations is fixed. Note that different values of parameter a can be chosen.

This method can also be applied to partial correlations sampled from different distributions. In order to do this we just need to replace the equation in 3.2 with the appropriate expression for the expectation. A requirement however, is that the moments have closed form expressions.

The recursion used to compute the appropriate Beta parameters was implemented by Joe and Kurowicka [10] and we applied this to generate correlation matrices with fixed expectations.

3.2.1. Empirical properties of correlation matrices fixed expectations

In this subsection, we investigate the behaviour of correlation matrices generated when fixing the expectation of off-diagonals. Figures 3.5 and 3.6 show 3D scatter plots and marginal histograms for 1000 randomly generated (with permuted elements) 3×3matrices with fixed expectation, $\mu \in \{-0.4, 0.0, 0.4, 0.8\}$. The Beta parameters ($a = 1$, b_1 and b_2) we use to fix the expectation of the off-diagonals are given in the captions of Figures 3.5.

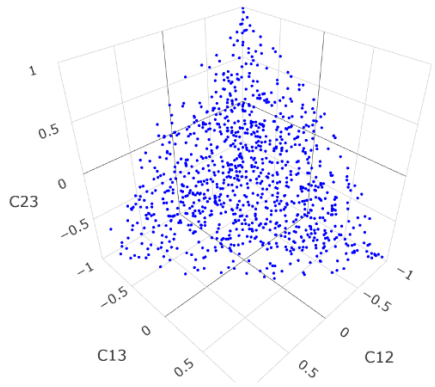
- At $\mu = 0.0$, the scatter plot 3.5b shows that the cloud of points fills the elliptical tetrahedron evenly. This is consistent with the fact that to achieve a partial correlation 0.0 we sample partial correlations from Beta(1,1). Therefore the partial correlations are symmetric around zero leading to approximately uniform coverage of the feasible region. The histogram 3.6b is flat , consistent with the uniform behaviour of the scatter plots.

- For $\mu = -0.4$ the partial correlations are sampled from Beta distributions $b_1 = 2.333$ and $b_2 = 34.764$. The points in the scatter plot are found in the lower region of the feasible space. The shape observed in 3.6a is skewed but due to low a and b_1 we still see variation in the values.

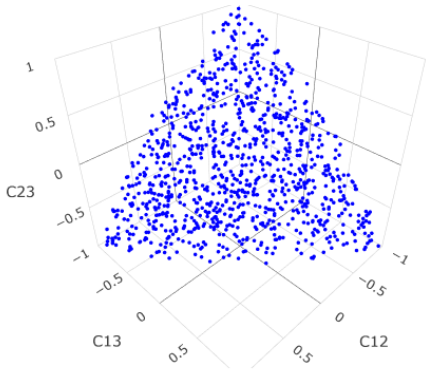
- For $\mu = 0.4$ we have that $b_1 = 0.429$ and $b_2 = 0.243$. We see in Figure 3.5c that the points begin to cluster toward the upper region of the elliptical tetrahedron, consistent with the positive target expectation. Due to the small values of b_1 and b_2 we see a strong skew in the histograms in Figure 3.6a, so while choosing a small a allows for variability, when μ becomes bigger the b values dominate to cause pronounced skewed distributions.

- When $\mu = 0.8$ $b_1 = 0.111$ and $b_2 = 0.001$. The points cluster in the top corner to achieve the target expectation, Figure 3.6d show highly skewed reflecting the extremely concentrated partial correlations. We see some matrices distributed along narrow ridges, this is due to the small value for a, as smaller values can still be attained.

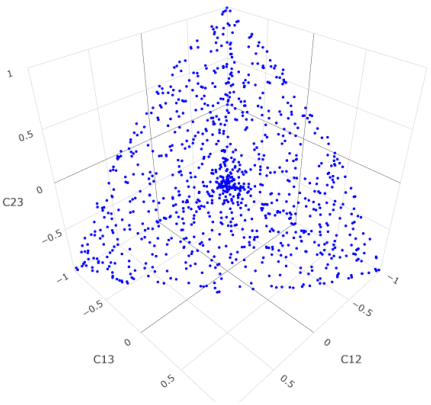
The difference in the shape distribution in the scatter plots 3.5 for low μ and high μ can be attributed to the geometry of the space of positive semidefinite matrices. For high correlations the region narrows and collapses to a sharp corner. While for lower correlations the shape exhibits a curved surface and no corner, seen in Figure 1.1



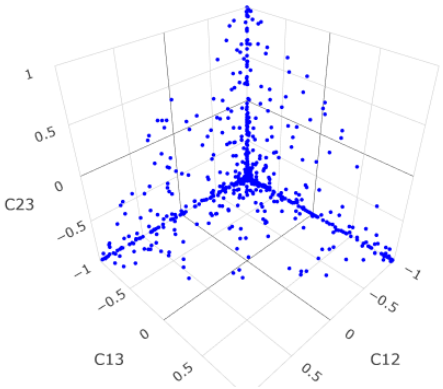
(a) 3D plot of samples $\mu = -0.4$, $a = 1, b_1 = 2.333, b_2 = 34.764$



(b) 3D plot of samples $\mu = 0.0$, $a = 1, b_1 = 1, b_2 = 1$

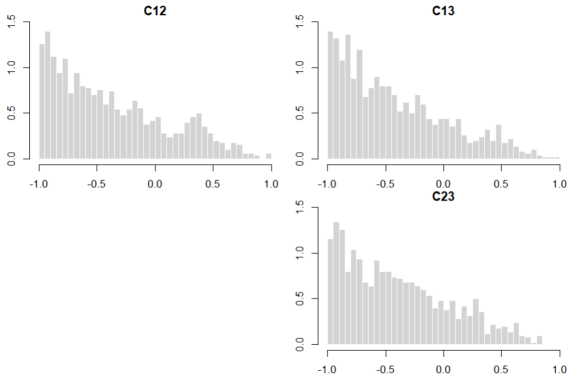


(c) 3D plot of samples $\mu = 0.4$, $a = 1$, $b_1 = 0.429, b_2 = 0.243$

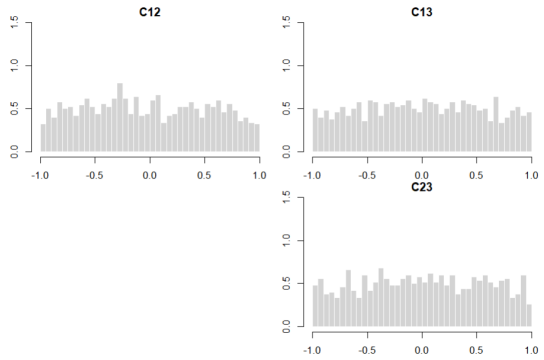


(d) 3D plot of samples $\mu = 0.8$, $a = 1, b_1 = 0.111, b_2 = 0.001$

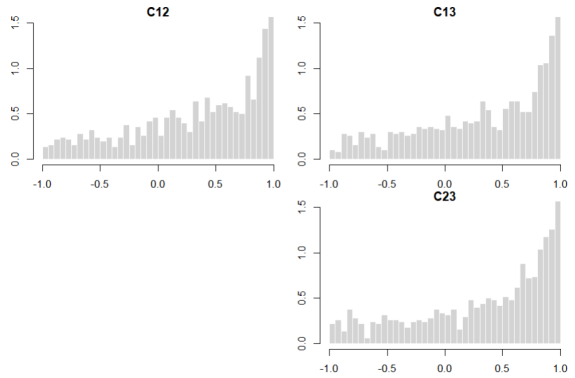
Figure 3.5: 3D scatter plots of sampled 3×3 permuted correlation matrices for different target expectations, $a = 1$.



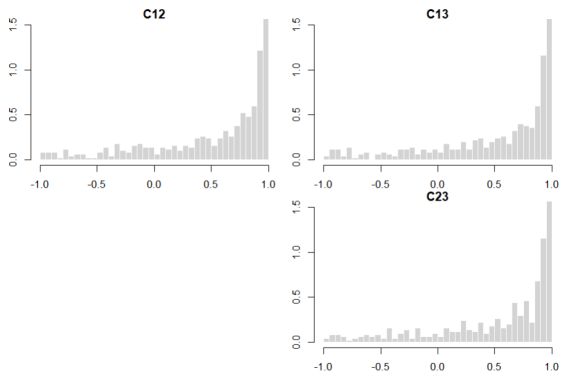
(a) Marginal histograms $\mu = -0.4$, $a = 1, b_1 = 2.333, b_2 = 34.764$



(b) Marginal histograms $\mu = 0.0$, $a = 1, b_1 = 1, b_2 = 1$



(c) Marginal histograms $\mu = 0.4$, $a = 1$, $b_1 = 0.429, b_2 = 0.243$



(d) Marginal histograms $\mu = 0.8$, $a = 1, b_1 = 0.111, b_2 = 0.001$

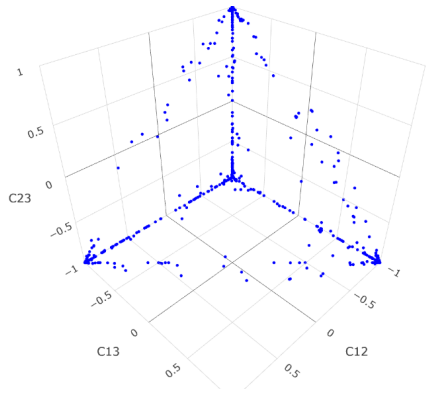
Figure 3.6: Marginal histograms of off-diagonal entries for sampled 3×3 permuted correlation matrices across different target expectations, $a = 1$.

So we have seen that if a is small (and b adjusted to match the target expectation), the variance increases, resulting in a wider spread of partial correlations. The greater variability hence leads to a wider distribution of the final correlation matrix entries.
Now we will look at scatter plots and marginal distributions where we keep μ the same but increase a. what we expect is when a increases, the Beta distribution becomes more concentrated around its mean and produces correlation matrices that cluster more tightly around the target expectation. In Figures 3.7 and 3.8 we take $\mu = 0.4$ and examine the behaviour for $a \in \{0.1, 1, 2, 10\}$.

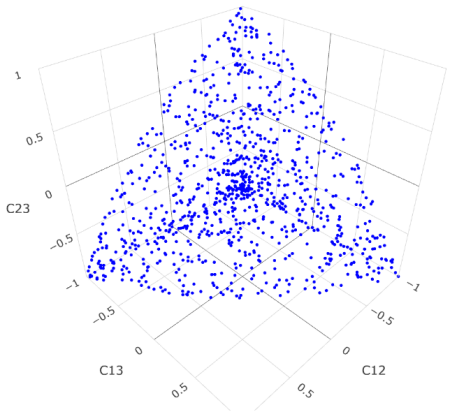
- When $a = 0.1$, the Beta distribution used is extremely U-shaped, with both a and b_k near zero. This causes most partial correlations to be sampled close to ± 1, resulting in full correlation matrices whose entries are pushed toward the boundaries of the feasible space. Ridges and corner clustering,can be seen in Figure 3.7a , and the marginal distributions of the correlation entries are distinctly bimodal, with peaks at -1 and $+1$.

- When $a = 1$, we have Beta(1,)the Beta distribution is skewed toward 1. This biases the partial correlations toward high positive values but allows for greater variability. The resulting correlation matrices are more spread out, with marginals that are peaked near $+1$ but no longer strictly bimodal.

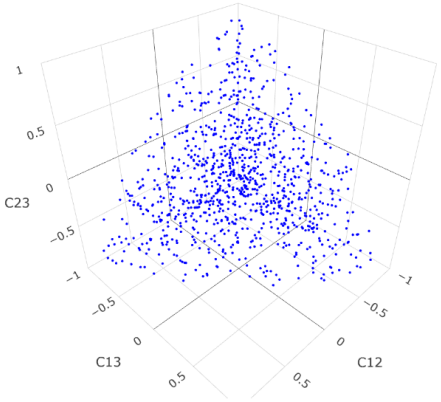
- Finally, for $a = 10$, the Beta distribution becomes concentrated around the target mean. The partial correlations in this case are tightly centred near 0.4, and the resulting full correlation matrices form a concentrated cluster around the matrix for which all off-diagonals are equal to 0.4. The marginals become smooth, bell-shaped distributions, and the 3D scatter plot no longer exhibits geometric ridges or boundary effects.



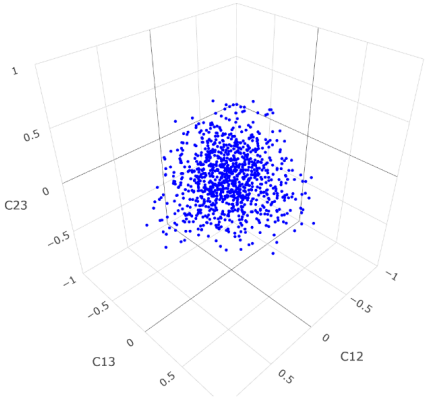
(a) 3D plot of samples $\mu = 0.4$, $a = 0.1$, $b_1 = 0.04, b_2 = 0.0014$



(b) 3D plot of samples $\mu = 0.4$, $a = 1$, $b_1 = 0.429, b_2 = 0.243$

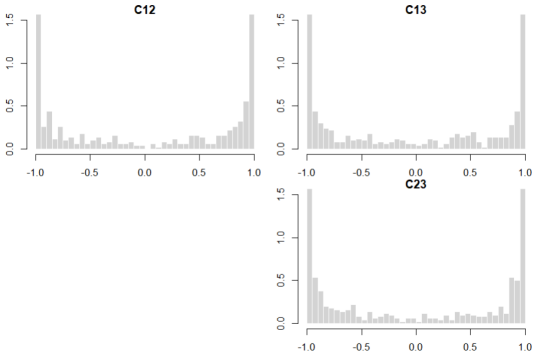


(c) 3D plot of samples $\mu = 0.4$ $a = 2$, $b_1 = 0.857$, $b_2 = 0.802$

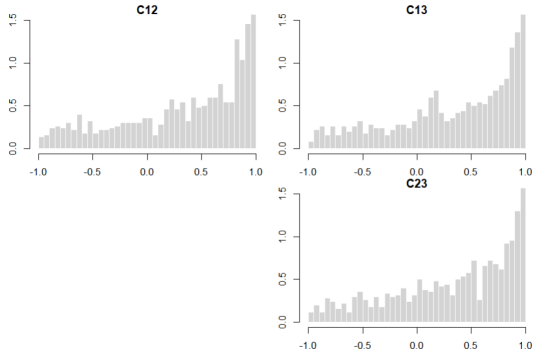


(d) 3D plot of samples $\mu = 0.4$, $a = 10$, $b_1 = 4.286$, $b_2 = 5.261$

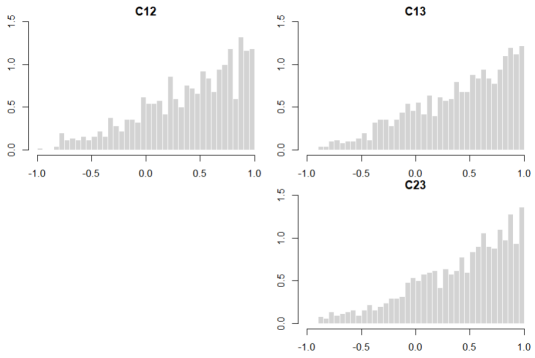
Figure 3.7: 3D scatter plots of sampled 3×3 permuted correlation matrices with expectation 0.4 for different values of parameter a.



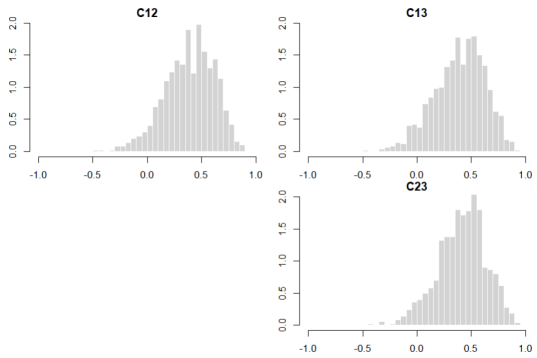
(a) Marginal histograms $\mu = 0.4$, $a = 0.1$, $b_1 = 0.04, b_2 = 0.001$



(b) Marginal histograms $a = 1$, $b_1 = 0.429, b_2 = 0.243$



(c) Marginal histograms $a = 2$, $b_1 = 0.857$, $b_2 = 0.802$



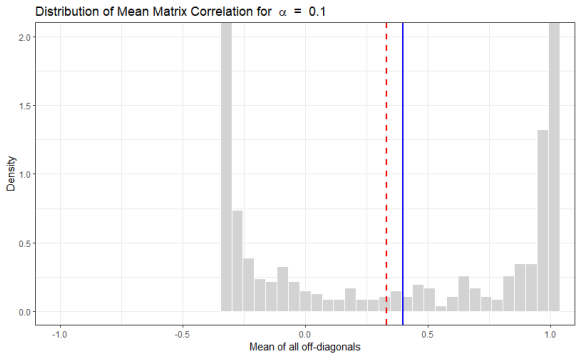
(d) Marginal histograms $a = 10$, $b_1 = 4.286$, $b_2 = 5.261$

Figure 3.8: Marginal histograms of off-diagonal entries for sampled 3×3 permuted correlation matrices with expectation 0.4 for different values of parameter a.

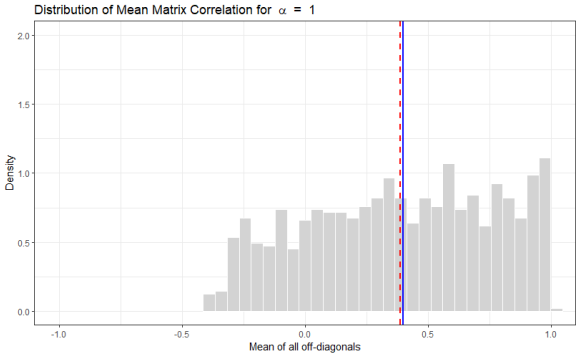
3.2.2. Distribution of average correlation

Now we investigate the distribution of average correlation in the correlation matrix. Figure 3.9 shows the histograms of the average correlation computed from 1000 simulations. Across samples, the expected value of off-diagonal elements is fixed, $\mu = 0.4$, and a is varied across $\{0.1, 1, 2, 10\}$ Let $\hat{\mu}$ denote the sample average correlation and μ still the target correlation. Each subplot contains red dotted line, indicating the sample average correlation $\hat{\mu}$ and a blue line indicating the target correlation, μ.

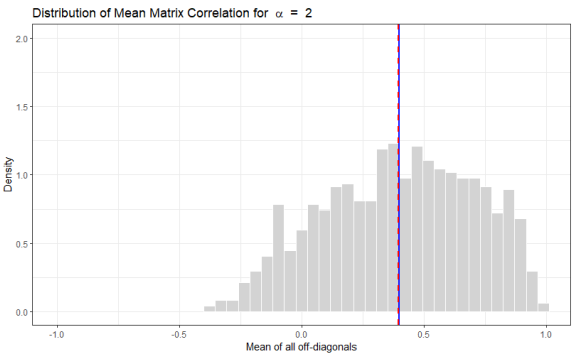
- For $a = 0.1$ we have that $\hat{\mu} = 0.334$. The distribution is extremely wide and U-shaped. This reflects the distribution of the margins, which are also concentrated at the boundaries (Figure 3.8a) . Therefore, in this case, the average correlation is lower than the target because the distribution is very broad.
- When $a = 1$, $\hat{\mu} = 0.387364$. The histogram in Figure 3.9b is now more concentrated around the target average correlation. The marginals are beginning to concentrate around the target expectation, but variance still causes some mismatch between μ and $\hat{\mu}$.
- For $a = 2$ and $a = 10$, the sampled average correlation and target essentially coincide, and the distribution of the averages becomes sharply peaked. Which makes sense as this parallels the behaviour of the marginal histograms in Figure 3.8.



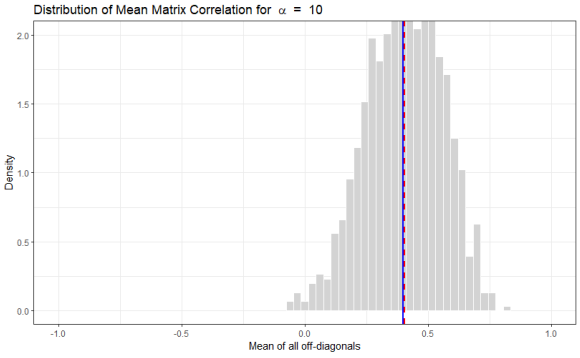
(a) $a = 0.1$ $\hat{\mu} = 0.3336023$



(b) $a = 1$ $\hat{\mu} = 0.387364$



(c) $a = 2$ $\hat{\mu} = 0.3967276$



(d) $a = 10$, $\hat{\mu} = 0.4034645$

Figure 3.9: Distribution of the average matrix correlation for $a = \{0.1, 1, 2, 10\}$

The expected value of the average correlation ($\hat{\mu}$) for $n = 3$ is given by:

$$\mathbb{E}[\bar{C}] = \frac{2}{3}\mu_1 + \frac{1}{3}\left(\mu_2 \cdot \gamma^2 + \mu_1^2\right).$$

As we saw in the Figures 3.9, the algorithm sets b_k based on μ and a to be such that the expectation of the marginals are a fixed, but for low a the variance is high. As a result we saw that the average correlation will not necessarily be the value that we fixed the expectations to. The entries are not symmetric, since C_{23} is constructed as a non-linear function of the partial correlations. With the permutations, each C_{ij} is identically distributed with probability $\frac{2}{3}$ of being distributed as the first level correlation and $\frac{1}{3}$ second level. The non-linear term $\mu_2 \cdot \gamma^2 + \mu_1^2$ is generally smaller than μ_1 when a is small. This has an influence on the expectation of the off-diagonals together. As a becomes bigger the beta distribution becomes increasingly concentrated. The variance becomes smaller and the partials will be more concentrated around the expectation. As a result of this $\mu_2 \approx \mu_1$ and $\gamma \approx \sqrt{1-\mu_1^2}$ and so

$$\mu_2\gamma^2 + \mu_1^2 \to \mu_1$$

this leads to $\mathbb{E}[\bar{C}] \to \mu_1$.The matrix becomes nearly deterministic, the matrix will concentrate around the matrix where all the off-diagonals are equal to μ_1.

In this chapter, we explored the construction of correlation matrices through the partial correlation C-vine parametrization. By leveraging the recursive structure of the C-vine, we showed how valid correlation matrices can be efficiently generated by sampling independent partial correlations from suitable distributions. We demonstrated how the marginals behave for both symmetric and asymmetric choices of the partial correlation distributions. Furthermore, we discussed recent extensions that allow for direct control over the expectation of the off-diagonal elements. Together, these results illustrate the power and adaptability of the C-vine partial correlation framework for both theoretical analysis and practical modelling of random correlation matrices.

4

Summary and discussion

The goal of this thesis was to compare two structured methods for generating correlation matrices, one based on SRD parametrization with global constraints, and one based on partial correlations C-vine parametrization. The aim was to understand how each method controls global properties, marginal structure. We examined both unconstrained and extended variants, including the approach by Tuitman et al. [14], which imposes a fixed average correlation through geometric constraints, and the extension by Joe and Kurowicka [10], which samples partial correlations from asymmetric beta distributions to fix the expected value of off-diagonal elements. Our comparison focused on theoretical construction, statistical behaviour and practical implementation, highlighting the trade-off between global control and local flexibility.

4.1. Theoretical foundations and constraints

Each method defines a different parametrization of the space of valid correlation matrices and imposes different types of constraints- ranging from strict geometric conditions to recursive probabilistic structure.

Chapter 2 begins by looking at the SRD parametrization of the correlation matrix, $C = TT'$, where T is a matrix whose rows t_i are independently sampled unit vectors. This method guarantees positive definiteness by construction and explores the space of correlation matrices uniformly given that no additional constraints are applied. The extension by Tuitman et al. [14] adds a global geometric constraint, the weighted sum of the vectors t_i must lie on a sphere of fixed radius, determined by the target average correlation ρ. This constraint forces later vectors in the sequence to lie in a shrinking feasible region resulting in increasingly restricted sampling space. The method guarantees that each matrix satisfies the average correlation exactly, but this control comes at the cost of flexibility, particularly in higher dimensions or for extreme target values of ρ.

The C-vine method takes a fundamentally different approach, using a recursive construction based on partial correlations $C_{ij;1,\ldots,k}$. These parameters are algebraically independent, any set of values $(-1,1)$ yields a valid, positive definite matrix through the recursive formula introduced in Joe [8]. Unlike the SRD parametrization, we do have some statistical control over the shape of the distribution of the matrices through the distribution of partial correlations chosen. Therefore although there is no explicit global constraint, like in the Tuitman et al. [14] method, the C-vine does allow for parametric control over the the matrices through the selection of distribution of the partial correlations.

An extension of the C-vine method allows the user to fix the expected value of each off-diagonal entry by choosing asymmetric Beta distributions for the partial correlations, with specific parameters. The algebraic independence of the parameters remains intact, but the distributions are no longer freely chosen: they must be coordinated to achieve the desired statistical behaviour. Unlike Tuitman et al. [14] method, this does not impose a hard constraint on the average correlation but rather concentrates the distribution around a chosen target mean μ. This means that while in Tuitman et al. [14] we find matrices that lie exactly on the plane of the desired average correlation, in this case we find matrices that concentrate around around the target value in expectation.

The Tuitman method enforces strict global control at the cost of parameter dependence and constrained sampling. The C-vine approach offers recursive, local flexibility and positive definiteness by design, with control over distributional behaviour through the choice of distribution for partial correlations.

4.2. Statistical properties of the matrices

The statistical properties provide an insight into how each method explores the space of valid correlation matrices and how constraints affect concentration and structure.

The C-vine method we can select specific beta distributions such that we achieve certain properties. For example if we choose Beta distribution with parameter $a = b = \alpha + \frac{n-k-1}{2}$ we achieve the LKJ distribution, for which the matrices generated are distributed proportional to $\det(C)^{\alpha-1}$. In addition, we can choose partial correlations sampled from asymmetric beta distributions with specifically chosen beta distributions such that the expectations of off-diagonals are fixed.

We begin by comparing unconstrained SRD parametrization and the C-vine method with LKJ parameters with $\alpha = 1$. In both cases the matrices are sampled uniformly from the space of valid correlation matrices. In the $n = 3$ case, this behaviour was confirmed in Figures 2.1 and 3.1 for both methods.

When we add the constraints from the Tuitman et al. [14] method, the matrix is forced to lie on a hyperplane of a constant average correlation. This corresponds to a slices through the full feasible region as seen in Figures 2.3. In contrast, the C-vine method fixes expectations by choosing specific beta distributions for partial correlations. By adjusting the parameter a, the distribution can be concentrated around matrices whose off-diagonal entries have expectation approximately equal to the target value μ, as illustrated in Figure 3.8.

The marginal distributions also display different properties. The unconstrained $C = TT'$ method show uniformly distributed marginals, Figure 2.1c, which make sense given that the vectors t_i are sampled uniformly. On the other hand, the C-vine parametrization with LKJ(1) distribution sample matrices uniformly from the feasible region but have Beta($\frac{n}{2}, \frac{n}{2}$) marginals, Figure 3.1c. Hence these two situations show similar distribution withtin the space of valid correlation matrices but have different underlying marginal distributions.

In the Tuitman et al. [14] method, imposing a fixed average ρ compresses the marginal distributions. The distribution remains unimodal but it more tightly centred around ρ. In contrast, the C-vine method with fixed expectation μ show Beta distributed marginals, while these marginals do not have the same beta distributions as the partials, they are similar.

Hence each method navigates the geometry of the space differently:

- Unconstrained methods: full-volume sampling
- Tuitman: plane, fixed constant average
- LKJ: tunable concentration near identity
- C-vine with fixed μ: tunable concentration near matrix with off-diagonals μ

4.3. Numerical Stability, Implementation, and Control

Beyond the mathematical properties there are important differences in terms of numerical stability, ease of implementation and level of control they offer over the matrices.

Both methods are computationally efficient, and scale well with matrix dimension n. This can be seen in Table 2.3 for the Tuitman et al method and in Table 1 of chapter 4 in Lewandowski et al. [11] for the C-vine parametrization.

The Tuitman et al extension has progressively tighter geometric constraints, which reduces numerical robustness when the target correlation ρ is high or the dimension is large. In such cases, the feasible region for the final vectors becomes small, increasing sensitivity to floating-point precision. The C-vine method, in contrast, constructs the matrix via recursive evaluation of closed-form expressions involving previously sampled partial correlations. This construction is numerically stable and does not suffer from accumulation of rounding error, like the Tuitman et al algorithm. When asymmetric Beta distributions are used (to fix expected correlations), extreme values of the partial correlations can occasionally lead to near-singular matrices, leading to reduced numerical stability.

The unconstrained $C = TT'$ parametrization is conceptually simple and easy to implement. The Tuitman et al extension requires more caution to implement, as mentioned small numerical inconsistencies can arise during the construction hence numerical safeguards must be considered, for example, to ensure values within square roots were not negative. The C-vine method has a more complex recursive structure, but once implemented, offers great flexibility. Partial correlations can be sampled independently, and the method naturally supports sampling from a wide range of distributions. When targeting specific expectations, the required parameters for the Beta distribution must be approximated numerically.

The primary trade-off between the methods is control versus flexibility. The Tuitman et al method offers strict control over the average correlation. Every matrix lies exactly on the hyperplane defined by the target ρ. However, this control introduces dependence between parameters and reduces sampling flexibility. In contrast, the C-vine method offers full local freedom in selecting the partial correlations. When using specific distribution families, such as symmetric Beta distributions with LKJ-parameters, one can control the overall structure of the generated matrices. With the extension, one can approximate a target average correlation across samples, but not enforce it explicitly per matrix. This results in greater distributional flexibility.

This comparative study clarifies how two approaches relate in terms of structure, sampling behaviour, and practical constraints. By analysing their theoretical properties and empirical output, we demonstrate that while the methods can yield similar distributions in the uniform case, their constrained versions lead to fundamentally different geometries and marginal behaviours. The Tuitman et al method is ideal when one needs to enforce a specific global structure, such as in stress testing or simulations where a fixed average correlation is required. The C-vine method is preferable when one needs statistical control over variability and marginal behaviour, such as in Bayesian modelling.

4.4. Discussion

While this thesis provides a comparative analysis of two structured correlation matrix generation methods, several theoretical and practical questions remain open.

One limitation of this study is that the analysis of Tuitman et al. [14] was conducted under the assumption of equal weighting, while the method was presented with these weights. This assumption simplifies the theoretical analysis and visualization. Extending the current analysis to incorporate this specific weights and investigating what the geometric implications are of having these could make this more realistic.

Furthermore, while the analysis of the C-vine partial correlation parametrization in this thesis was restricted to $n = 3$, the SRD parametrization was also explored for higher dimensions. A next step would be to look at higher dimensions in the C-vine setting and look at the marginal distributions, however due to time constraints, this was not presented in this thesis.

In hindsight these methods differ in aims. The Tuitman approach emphasizes hard constraints and geometric feasibility, making it well suited to stress testing and worst-case analysis. The C-vine method, by contrast aligns with a Bayesian goal, where variability and tunable prior is important. Hence the appropriate method depends on whether the user wants control over what is generated or flexibility in what is likely. This makes these methods fundamentally different.

One potential extension could be to relax the fixed average correlation in the Tuitman et al method. Suppose we would like the average correlation to fall within an interval, a solution is to generate matrices for a range of ρ and aggregate the samples, effectively stacking the hyperplanes in the correlation space. This is illustrated in Figure4.1, where 10 values of ρ were uniformly sampled from the interval $\rho \in [0.15, 0.25]$, and for each value, 100 correlation matrices were generated. The resulting slices were then stacked to form a thick disk. While this approach works in principle, it is not efficient and lacks theoretical elegance. A more integrated solution might involve modifying the norm constraint to allow variability in the final norm s. However, this introduces the risk of sampling infeasible vector lengths at intermediate steps or violating positive definiteness in the final matrix. Hence developing a consistent and feasible algorithm that produces matrices with average correlation in a desired interval is an interesting challenge.

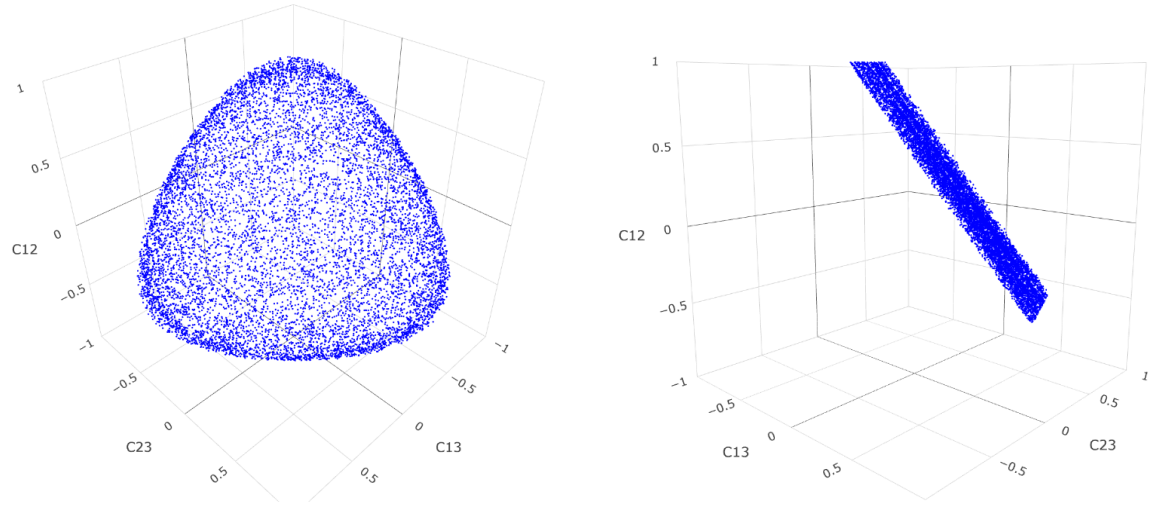


Figure 4.1: 1000 matrices generated with average correlation $\rho \in [0.15, 0.25]$

In the current construction by Tuitman et al, the lengths l_i of the vectors t_i are sampled uniformly from their feasible intervals. Modifying this sampling method would change the distribution of the resulting correlation matrices along the hyperplane. Alternative sampling strategies such as biased length distribution will affect the geometry and marginal behaviour of the generated matrices. Additionally one could modify the sampling of the first vector t_1, either by drawing it from a specific distribution or by constraining its direction. Since all subsequent vector are conditioned on earlier ones, such as change would propagate through the construction and likely alter both the feasible region and the final d istribution. Analysing these dependencies could offer further insight into how local sampling choices influence global matrix properties.

Another direction concerns the marginal structure of matrices for $n > 3$ for the Tuitman method. The marginal distributions of the off-diagonal entries exhibit asymmetry, as shown in Figure 2.6. This asymmetry arises from the fixed ordering of the vectors during construction. A straightforward remedy would be to apply random permutations to the matrix elements. It would then be interesting to study the resulting permuted distribution more closely to investigate whether known distributions, such as transformed Beta, can approximate these marginals analytically.

Finally, while this thesis focused on the average correlation, a natural extension is to consider other global matrix statistics, such as the variance of off-diagonal elements. These quantities are non-linear functions of the matrix entries and therefore are more difficult to control directly. In the C-vine construction it is currently possible to influence the variance of the off-diagonal entries indirectly by adjusting the parameters of the partial correlation distributions. Similarly one can control the variability in the average correlation across matrices in expectation. However it would be an interesting direction to investigate whether the variance can be controlled more directly through a recursive formulation rather than distributional tuning.

All of the figures and data presented in this thesis was computed in R-studio. For access to the implementations or underlying code.

Bibliography

[1] Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31 of *MSRI Publications*, pages 1–58. Cambridge University Press, 1997. URL https://math.uchicago.edu/~shmuel/AAT-readings/Combinatorial%20Geometry,%20Concentration,%20Real%20Algebraic%20Geometry/ball.pdf.

[2] Tim Bedford and Roger Cooke. Vines - a new graphical model for dependent random variables. *Annals of Statistics*, 30, 09 1999. doi: 10.1214/aos/1031689016.

[3] Carole Bernard, Ludger Rüschendorf, and Steven Vanduffel. Value-at-risk bounds with variance constraints. *Journal of Risk and Insurance*, 84(3):923–959, 2017. doi: https://doi.org/10.1111/jori.12108. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jori.12108.

[4] Phelim Boyle and Thierno Bocar N'Diaye. Correlation matrices with the perron frobenius property. *Electronic Journal of Linear Algebra*, 34:240–268, 06 2018. doi: 10.13001/1081-3810.3616.

[5] Kian Ming A. Chai. Three-by-three correlation matrices: its exact shape and a family of distributions. *Linear Algebra and its Applications*, 458:589–604, 2014. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2014.06.039. URL https://www.sciencedirect.com/science/article/pii/S002437951400408X.

[6] Alvaro Jóse Flórez, Ariel Alonso Abad, Geert Molenberghs, and Wim Van Der Elst. Generating random correlation matrices with fixed values: An application to the evaluation of multivariate surrogate endpoints. *Computational Statistics Data Analysis*, 142:106834, 2020. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2019.106834. URL https://www.sciencedirect.com/science/article/pii/S0167947319301896.

[7] Amelie Hüttner and Jan-Frederik Mai. Simulating realistic correlation matrices for financial applications: correlation matrices with the perron–frobenius property. *Journal of Statistical Computation and Simulation*, 89:1–22, 11 2018. doi: 10.1080/00949655.2018.1546861.

[8] Harry Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177 – 2189, 2006. doi: 10.1016/j.jmva.2005.05.010. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-33749138913&doi=10.1016%2fj.jmva.2005.05.010&partnerID=40&md5=7a7957a8f97c473619c27620dc235543. Cited by: 169; All Open Access, Bronze Open Access.

[9] Harry Joe. Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis*, 97(10):2177–2189, 2006. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2005.05.010. URL https://www.sciencedirect.com/science/article/pii/S0047259X05000886.

[10] Harry Joe and Dorota Kurowicka. Random correlation matrices generated via partial correlation c-vines. Preprint submitted to the Journal of Multivariate Statistics, 2025.

[11] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2009.04.008. URL https://www.sciencedirect.com/science/article/pii/S0047259X09000876.

[12] George Marsaglia and Ingram Olkin. Generating correlation matrices. *SIAM Journal on Scientific and Statistical Computing*, 5(2):470–475, 1984. doi: 10.1137/0905034. URL https://doi.org/10.1137/0905034.

[13] Peter J. Rousseeuw and Geert Molenberghs. The shape of correlation matrices. *The American Statistician*, 48(4):276–279, 1994. ISSN 00031305. URL http://www.jstor.org/stable/2684832.

[14] Jan Tuitman, Steven Vanduffel, and Jing Yao. Correlation matrices with average constraints. *Statistics Probability Letters*, 165:108868, 2020. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2020.108868. URL https://www.sciencedirect.com/science/article/pii/S0167715220301711.

[15] Ruodu Wang, Liang Peng, and Jingping Yang. Bounds for the sum of dependent risks and worst value-at-risk with monotone marginal densities. *Finance and Stochastics*, 17(2):395–417, 2013. ISSN 1432-1122. doi: 10.1007/s00780-012-0200-5. URL https://doi.org/10.1007/s00780-012-0200-5.

[16] Hongwei Yang, Wing Hang Wong, Kelly Bradley, and Michael Toland. Partial and Semi-Partial Correlations for Categorical variables in Educational Research: Addressing two common misconceptions. *General Linear Model Journal*, 43(1):1–15, 1 2017. doi: 10.31523/glmj.043001.001. URL https://doi.org/10.31523/glmj.043001.001.

A

Appendix A

c_{ij}	c_{kl}	Corr.
c_{12}	c_{13}	−0.154
c_{12}	c_{14}	0.134
c_{12}	c_{15}	0.126
c_{12}	c_{23}	−0.167
c_{12}	c_{24}	0.167
c_{12}	c_{25}	0.130
c_{12}	c_{34}	−0.681
c_{12}	c_{35}	−0.707
c_{12}	c_{45}	−0.572
c_{13}	c_{14}	0.041
c_{13}	c_{15}	0.028
c_{13}	c_{23}	−0.035
c_{13}	c_{24}	−0.488
c_{13}	c_{25}	−0.497
c_{13}	c_{34}	0.114
c_{13}	c_{35}	0.156
c_{13}	c_{45}	−0.333
c_{14}	c_{15}	0.160
c_{14}	c_{23}	−0.484
c_{14}	c_{24}	−0.096
c_{14}	c_{25}	−0.312
c_{14}	c_{34}	−0.053
c_{14}	c_{35}	−0.318
c_{14}	c_{45}	−0.012
c_{15}	c_{23}	−0.513
c_{15}	c_{24}	−0.311
c_{15}	c_{25}	−0.122
c_{15}	c_{34}	−0.339
c_{15}	c_{35}	−0.095
c_{15}	c_{45}	0.143
c_{23}	c_{24}	0.051
c_{23}	c_{25}	0.032
c_{23}	c_{34}	0.144
c_{23}	c_{35}	0.128
c_{23}	c_{45}	−0.323
c_{24}	c_{25}	0.161
c_{24}	c_{34}	−0.064
c_{24}	c_{35}	−0.389
c_{24}	c_{45}	−0.020
c_{25}	c_{34}	−0.352
c_{25}	c_{35}	−0.111
c_{25}	c_{45}	0.125
c_{34}	c_{35}	0.399
c_{34}	c_{45}	0.271
c_{35}	c_{45}	0.395

Table A.1: Correlations between unique off-diagonal elements of a 5×5 correlation matrix.