*Journal of Statistical Software*

*April 2022, Volume 102, Issue 3.* *doi: 10.18637/jss.v102.i03*

# covsim: An R Package for Simulating Non-Normal Data for Structural Equation Models Using Copulas

**Steffen Grønneberg**
BI Norwegian Business School

**Njål Foldnes**
BI Norwegian Business School

**Katerina M. Marcoulides**
University of Minnesota

**Abstract**

In factor analysis and structural equation modeling non-normal data simulation is traditionally performed by specifying univariate skewness and kurtosis together with the target covariance matrix. However, this leaves little control over the univariate distributions and the multivariate copula of the simulated vector. In this paper we explain how a more flexible simulation method called vine-to-anything (VITA) may be obtained from copula-based techniques, as implemented in a new R package, **covsim**. VITA is based on the concept of a regular vine, where bivariate copulas are coupled together into a full multivariate copula. We illustrate how to simulate continuous and ordinal data for covariance modeling, and how to use the new package **discnorm** to test for underlying normality in ordinal data. An introduction to copula and vine simulation is provided in the appendix.

*Keywords*: non-normal simulation, covariance model, vine copulas, ordinal covariance models, R.

## 1. Introduction

Structural equation modeling (SEM) and factor analysis are regularly applied to data in the psychological, educational, business, behavioral, and medical sciences. The central component in these methods is the covariance matrix from which the model parameters are identified. In this article we present software for simulating from a class of distributions with a fixed covariance matrix, which therefore can be used in SEM simulation studies. This distributional class is more flexible than the methods currently in use, and may therefore extend the range of conditions investigated with simulations.

When the examined data are continuous, the most popular SEM estimation method used is the normal-theory based maximum likelihood (NTML) method (Jöreskog 1967) which is asymptotically efficient when data are normally distributed. NTML estimation is a special case of a class of moment based estimators known as minimum discrepancy function estimators (see, e.g., Shapiro 1983), and is therefore known to be consistent also under non-normality. When using classical standard errors with NTML, valid inference is attained mainly when the data are normal. While several standard error and test statistic formulas have been proposed in order to robustify inference with NTML and other minimum discrepancy estimators under non-normality (e.g., Satorra and Bentler 1988; Wu and Lin 2016; Marcoulides, Foldnes, and Grønneberg 2020), their performance depends heavily on the distribution and sample size of the data (e.g., Curran, West, and Finch 1996; Fouladi 2000; Foldnes and Olsson 2015; Grønneberg and Foldnes 2019b). In settings with ordered-categorical data, least squares estimation based on polychoric correlations is the most prevalent estimation method (Christoffersson 1977; Muthén 1984). Polychoric correlations are essentially the correlations among continuous bivariate normally distributed vectors underlying the observed ordinal data, but may be heavily biased outside normality (Foldnes and Grønneberg 2019b, 2022b). Hence, under both continuous and ordinal data analyses, the normality assumption is a central starting point for estimation and inference.

Unfortunately, in most empirical research situations data are seldom drawn from populations in which the normality assumption holds exactly (Micceri 1989; Cain, Zhang, and Yuan 2017). And while several estimation and inference methods that do not assume normality have been suggested (for an overview, see Tarka 2018), it is in most conditions not feasible to analytically derive results on their performance as a function of the data generating distribution. Monte Carlo simulation studies have as a result become essential tools for evaluating the behavior of various aspects of SEM techniques, such as parameter and standard error bias, performance of test statistics, and power calculations, relative to distributional characteristics (Boomsma 2013). The external validity of these studies is weakened if the chosen data generation mechanism does not resemble the real-world distributions encountered in the relevant field of practice. To be able to model such distributions, we need simulation methods that can match a given covariance matrix but still offer distributional flexibility.

The aims of the present paper are twofold. Firstly, to present the R (R Core Team 2021) package **covsim** (Foldnes and Grønneberg 2022a), available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=covsim`, which implements non-normal data simulation methods proposed by Foldnes and Olsson (2016) and Grønneberg and Foldnes (2017). Both methods generate data with a prescribed covariance matrix. Our emphasis will be on the method called vine-to-anything (VITA), since it offers greater flexibility that we deem particularly useful to SEM methodologists. For instance, the flexibility of VITA renders it uniquely well-suited for being employed in simulation studies with ordinal SEM, as further discussed in Section 6.2.

VITA is a simulation method based on vine copulas. Copulas are multivariate distributions with uniform marginals, and vine copulas are a special type of copulas. Since copulas, vines and multivariate simulation theory are not well known to SEM practitioners and methodologists, the second aim of the paper is to introduce these topics during our presentation of the VITA method and the **covsim** package. We also include a technical appendix with an elementary though mathematically complete introduction to multivariate simulation theory with vine copulas, as this seems to be missing from the literature.

We next give an overview of statistical software for drawing data from non-normal multivariate distributions with a predefined covariance matrix. The classical and still most frequently used approach is that of Vale and Maurelli (1983), where the user specifies the univariate skewness and kurtosis. This method is currently the only option in popular commercial software such as **EQS** (Bentler 2006) and **LISREL** (Jöreskog and Sörbom 2006), and in the widely used R package **lavaan** (Rosseel 2012). Other approaches that also focus on controlling moments have recently been proposed. The independent generator approach proposed by Foldnes and Olsson (2016) can match pre-specified univariate skewness and kurtosis, and is more flexible than the Vale-Maurelli method. This method is available in the `rIG()` function in package **covsim**, and its use is described in a later section. Recently Qu, Liu, and Zhang (2019) used independent generator variables in a method which controls multivariate skewness and kurtosis, at the expense of control over univariate skewness and kurtosis. This method is available in package **mnonr** (Qu and Zhang 2020). A method that fully controls the univariate distributions (not only the lower-order moments) is the NORTA method of Cario and Nelson (1997), which is implemented in package **SimCorMultRes** (Touloumis 2016). The Vale-Maurelli, independent generator and NORTA approaches have the great benefit of being technically easy to analyze and implement. For instance, the technical tractability allows the asymptotic covariance matrix of the empirical covariances to be exactly calculated (Foldnes and Grønneberg 2017). The simplicity of the methods also allows for fast simulation. However, the simplicity and speed of these methods come at a cost: NORTA always has a normal copula (Cario and Nelson 1997), while Vale-Maurelli in most cases has a normal copula (Foldnes and Grønneberg 2015). This means that the true multivariate dependence structure does not depart from that of the multivariate normal distribution. In addition, only NORTA completely controls the univariate marginals. To the best of our knowledge, besides the approach taken in the present article, there is only one method that offers some control of the copula when simulating from distributions with a given covariance matrix. Mair, Satorra, and Bentler (2012) proposed a two-stage data-generation process where a very large sample is first simulated from a copula combined with marginal specification, whose distribution we denote by $F_{\text{pre}}$. Then the inverse of a square root of the sample covariance matrix from this large sample is computed. To simulate data, in the second stage a sample of desired size is drawn from $F_{\text{pre}}$, and multiplied by first the inverse of the square root matrix from the previous stage and then by a square root matrix of the target covariance. The two-stage approach guarantees that the rows are independent and identically distributed, and it follows from construction that the simulated vector has the correct population covariance matrix. Also, the simulated vector has a non-normal copula, provided $F_{\text{pre}}$ was chosen to have a non-normal copula. However, both the margins and the copula are distorted by the post-multiplication of square root matrices. That is, although $F_{\text{pre}}$ is fully specified in terms of multivariate copula and univariate distributions, the simulated vector does not inherit this copula nor the margins and control is lost both in terms of copula and marginal distributions. Mair *et al.* (2012) illustrated their code using common multivariate copula families using Gumbel and Clayton copulas, as implemented in package **copula** (Hofert, Kojadinovic, Maechler, and Yan 2020). An implementation of this method is available in package **simsem** (Pornprasertmanit, Miller, Schoemann, and Jorgensen 2021). The flexibility of this implementation is presently limited to a rather restricted class of multivariate copulas, comprising elliptical, Archimedean, extreme-value and some other copula families available in package **copula**.

VITA improves upon the approach of Mair *et al.* (2012) by allowing complete control of the $p$ marginal distributions, of the bivariate copulas of a chosen set of $p-1$ pairs of variables, and of certain conditional bivariate copulas of the remaining $(p-1)(p-2)/2$ pairs of variables.

The increased degree of control and flexibility of our approach relative to existing methods is made possible by employing the powerful multivariate copula-based construction called a regular vine. A primary aim of the present article is to present and illustrate our approach using the newly developed **covsim** package.

The remainder of the article is organized in the following way. First, we explain the copula approach in a two-dimensional setting. We then demonstrate a very flexible copula-based approach to non-normal simulation in the two-dimensional case. An important advantage of this approach is that the copula class and the exact marginal distributions of the two-dimensional case may be fully specified by the user. Next, we develop the full multivariate extension, which still allows for complete control of the marginal distributions, and considerable flexibility in the dependence structure. We then detail the implementation of the **covsim** package, and end the paper with two additional examples: First, we show how to simulate from a non-normal continuous SEM with fixed parameters, and then we show how to simulate data for an ordinal SEM. Example code is provided throughout the paper, and complete replication code is available in the online supplementary material. The appendix provides an introduction to implementing the simulations on a computer, and follows the progression of the paper.

Throughout this article we illustrate the capacity of **covsim** in the context of simple structural equation modeling settings. We use only a limited set of non-normal distributional conditions in each illustration, and we caution that the external validity of our findings is therefore limited. To enhance the validity a larger range of distributional and sample size conditions must be included.

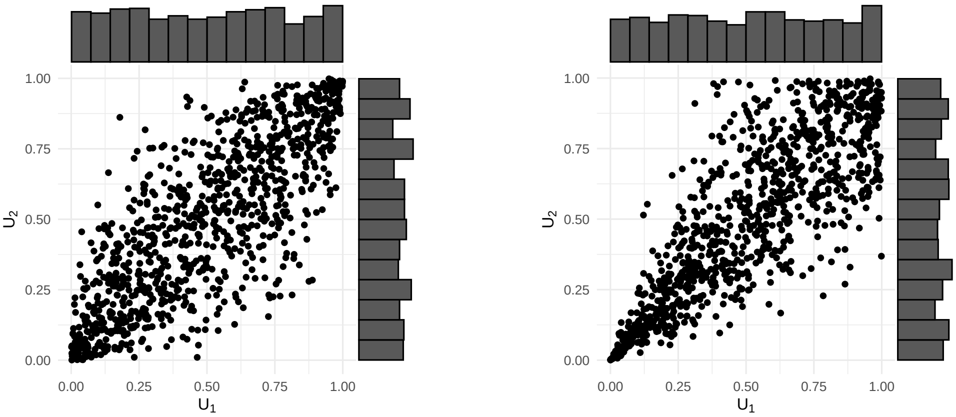
## 2. The bivariate case

We start by considering the bivariate case. Our aim is to introduce the concept of a copula and how it can be used to simulate non-normal random variables with a given correlation. In subsequent sections we extend our simulation procedure to the general multivariate case. For a textbook treatment of copula theory, see Nelsen (2007). Note that this book, together with most books on copulas, assume that the reader has a strong mathematical background, including some measure theory, and that we do not assume such a background in the current presentation. Some useful introductory papers on copulas which can be read without such a background are Yan (2007); Genest and Favre (2007); Frees and Valdez (1998).

A copula is a distribution with uniform univariate margins. Copulas are used to describe the dependency structure between variables, when taking the marginal distributions out of the equation. There are many classes of copulas, and within each class there is typically a parameter that controls the strength of dependence. We start with the normal copula. Let $\Phi(x)$ denote the cumulative distribution function (CDF) of a standard normal distribution, and let $\Phi^2(x, y; \rho)$ denote the CDF of the bivariate standard normal distribution with correlation parameter $\rho$, i.e., $\Phi^2(x, y; \rho) = P(Z_1 \leq x, Z_2 \leq y)$ where $Z_1, Z_2$ are bivariate normal and standardized, and have correlation $\rho$. Then the normal copula with parameter $\rho \in (-1, 1)$ is given by

$$C^N(u_1, u_2; \rho) = \Phi^2(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho).$$

As an example of a non-normal copula class, consider Clayton copulas, which are parameter-



(a) Normal copula with $\rho = 0.8$.

(b) Clayton copula with $\theta = 3.4$.

Figure 1: Random samples of size $n = 1000$ drawn from two bivariate copulas.

ized by the dependence parameter $\theta \in (0, \infty)$:

$$C^{Cl}(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta}.$$

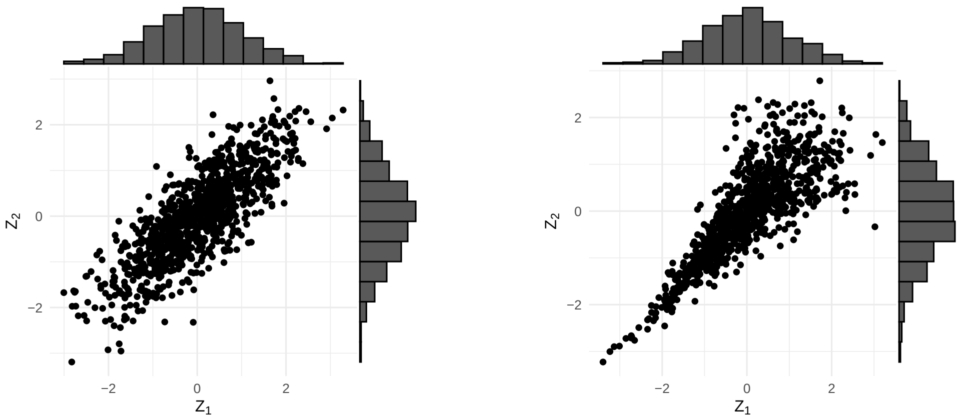
Clayton copulas are useful for modeling lower tail dependence, a measure of dependence between two variables in the lower left tail of the joint distribution. Figure 1 depicts random draws of size $n = 1000$ from each of these copulas. We set $\rho = 0.8$ for the normal copula and $\theta = 3.4$ for the Clayton copula. In both Figures 1a and 1b we see that the marginal empirical distributions are close to uniform. A notable difference is the lower tail dependence in Figure 1b which does not appear in Figure 1a.

Bivariate copulas are important since they constitute one of two fundamental building blocks for bivariate distributions. The other building block consists of the two univariate marginal distributions. A fundamental theorem (Sklar 1959) guarantees that any bivariate distribution may be decoupled into a bivariate copula and the two marginal distributions, and vice versa; given two marginal distributions $F_1(x_1)$ and $F_2(x_2)$ and a copula $C(u_1, u_2; \theta)$, then

$$F(x_1, x_2) := C(F_1(x_1), F_2(x_2); \theta) \tag{1}$$

is a valid bivariate CDF, whose univariate margins are distributed according to $F_1(x_1)$ and $F_2(x_2)$. For instance, if $F_1$ and $F_2$ are the standard normal distribution, the bivariate distributions stemming from the normal copula with $\rho = 0.8$, and from the Clayton copula with $\theta = 3.4$, will both result in bivariate distributions with standard normal marginals, and with a Pearson correlation of 0.8. That is, setting $\theta = 3.4$ yields a Clayton copula such that when combined with standard normal marginals will yield a distribution with $\rho = 0.8$. Figure 2 shows random samples from these two distributions, obtained by applying the standard normal quantile function to the observations in Figure 1, which will change the marginals of the simulated data to be standard normal. Figure 2 depicts two very different bivariate distributions. Although sharing the same standard normal marginal distributions, and the same correlation coefficient 0.8, it is clear that the distribution in Figure 2b is far from the bivariate normal distribution in Figure 2a.

This illustration hints at the following process for a researcher that wants to simulate data from a bivariate distribution with pre-specified covariance and univariate marginals:



(a) Normal distribution with $\rho = 0.8$.

(b) Clayton distribution with $\theta = 3.4$.

Figure 2: Random samples of size $n = 1000$ drawn from two bivariate distributions with standard normal marginals and correlation 0.8.

1. Specify marginal distributions $F_1$ and $F_2$ and specify a target covariance.
2. Specify a bivariate copula class $C(u_1, u_2; \theta)$, with dependence parameter $\theta$.
3. Use a numerical procedure to determine $\theta_0$ so that $C(F_1(x_1), F_2(x_2); \theta_0)$, the coupled distribution, has the pre-specified covariance.

For a given set of marginals, and a given copula, the set of attainable covariances is usually constrained. Then, in step 3 there is no solution $\theta_0$. In such a case, the copula class or the marginal specifications should be adjusted.

The three steps are conducted in the **covsim** package in R as follows.

```
R> library("covsim")
R> mnorm <- list(list(distr = "norm"), list(distr = "norm"))
R> sigma.target <- matrix(c(1, 0.8, 0.8, 1), 2)
R> set.seed(1)
R> calibrated.vita <- vita(mnorm, sigma.target, family_set = "clayton")

Tree 1
    1 - 2 ( 1 of 1 )

R> summary(calibrated.vita)

$margins
# A data.frame: 2 x 2
 margin distr
      1  norm
      2  norm

$copula
```

```
# A data.frame: 1 x 10
 tree edge conditioned conditioning var_types  family rotation parameters df
    1    1        1, 2                      c,c clayton        0        3.4  0
  tau
 0.63
```

We verify that the target covariance matrix has been attained, using the `rvine()` function from package **rvinecopulib** (Nagler and Vatter 2021). This package offers fast simulation from vines.

```
R> library("rvinecopulib")
R> cov(rvine(10^5, calibrated.vita))
```

```
          [,1]      [,2]
[1,] 0.9994719 0.8017954
[2,] 0.8017954 0.9985791
```

As indicated above, we may simulate from a distribution of the form $C(F_1(x_1), F_2(x_2); \theta)$ by first simulating $(U_1, U_2)$ from the copula $C$, and then apply the quantile functions of the marginals to each coordinate. That is, $(F_1^{-1}(U_1), F_2^{-1}(U_2))$ has marginals $F_1, F_2$ and copula $C$, meaning its full distribution equals $C(F_1(x_1), F_2(x_2); \theta)$. See the technical appendix for an explanation for why this is so and how to simulate from a copula.

## 3. The trivariate case: Introducing vines

In the previous section we studied bivariate copulas, and the calibration of their dependence parameter so that the coupling of given marginals will meet a target covariance. There are many classes of bivariate copulas, but few classes of higher-dimensional copulas. In this section we will circumvent the lack of parametric multivariate copula classes by using a statistical construction called a regular vine (Bedford and Cooke 2002). Vines allow us to construct multivariate copula distributions by combining two-dimensional copulas. For the purpose of covariance modeling and simulation, the procedure detailed here was originally proposed by Grønneberg and Foldnes (2017).

Let us first proceed to the case of three variables. Our goal is to construct distributions with given marginal univariate distributions for each of the three variables, and with a given $3 \times 3$ covariance matrix. Imagine a researcher is concerned with whether non-normal correlated errors in growth curve modeling may affect the quality of inference for the correlation $\rho$ between the intercept and slope factors. The default in growth curve modeling is to assume that residual errors are mutually independent across measurement occasions. However, correlated errors may be meaningful as they represent carryover effects from previous occasions not accounted for by the intercept and the linear slope latent variable (Grimm and Widaman 2010; Marcoulides 2019). The issue of how residual error structures in latent growth curve modeling should be specified (e.g., as constrained, free, independent, autocorrelated, homogeneous, or non-homogeneous) is currently of great concern in the SEM literature, as it is now becoming much more recognized that considerable bias in the latent variable variance-covariance matrix can arise from the improper specification of these errors (Dimitrov 2002;

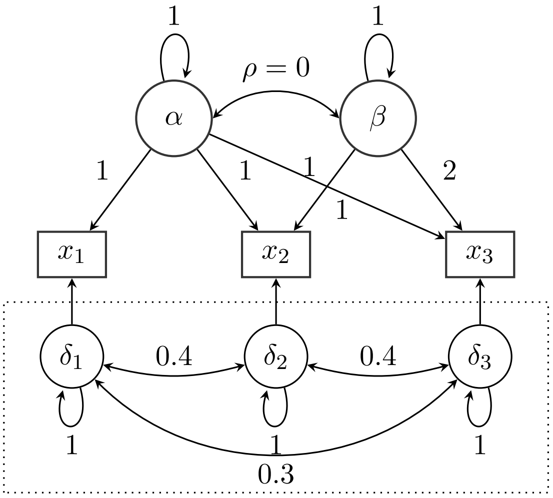


Figure 3: Linear growth curve population model with correlated residual errors.

Van De Schoot, Sijbrandij, Winter, Depaoli, and Vermunt 2017; Marcoulides 2019; Laenen, Alonso, Molenberghs, and Vangeneugden 2009; Grimm and Widaman 2010).

The researcher wants to simulate data with correlated residual errors and sets up a simple linear growth model, see Figure 3, where the population values are indicated: There is zero correlation $\rho$ between the slope and the intercept, all latent variables have unit variance, and the errors $\delta = (\delta_1, \delta_2, \delta_3)^\top$ are correlated with covariance matrix

$$\Sigma_\delta = \begin{pmatrix} 1 & 0.4 & 0.3 \\ 0.4 & 1 & 0.4 \\ 0.3 & 0.4 & 1 \end{pmatrix}.$$

The researcher is concerned with non-normality in the error vector $\delta$, and therefore wants to construct a trivariate non-normal distribution whose covariance is $\Sigma_\delta$.

We remark that the illustrations in the present article do not discuss the reasons behind specific choices of marginal distributions and dependence structure. Such choices depend on the purpose of the simulation study. Our aim in this and following illustrative analyses is simply to demonstrate how vine constructions work. However, we note that routines exist to select best-fitting vine structures and bicopula families relative to an existing real-world dataset (e.g., function `vinecop()` in package **rvinecopulib**). This could be done to increase external validity of simulation studies. For an example of how to construct a VITA distribution based on a well-known empirical dataset, see Grønneberg and Foldnes (2017, Section 3.2). General papers on the selection and usage of copulas are Yan (2007); Genest and Favre (2007); Embrechts, Lindskog, and Mcneil (2003); Grønneberg and Hjort (2014), an influential paper on vine-based modeling is Aas, Czado, Frigessi, and Bakken (2009), and a book with practical issues on vine modeling is Kurowicka and Joe (2011).

First, the researcher considers the three univariate error distributions, and decides that the first error should be standard normally distributed, the second error should be a scaled chi-squared distribution with one degree of freedom (DF), and the third error should follow a scaled Student's $t$ distribution with five DFs. The scalings are necessary to obtain unit variance in the two latter distributions. Clearly, it might be questioned whether the case of different error distributions for the same variable at different measurement occasions is realistic, but since our main purpose here is to illustrate the flexibility of VITA, we proceed with three different error distributions.

Even though the marginal distributions in $\delta$ have now been specified, there are still many trivariate distributions with these marginals and with covariance $\Sigma_\delta$. The specification will be complete once the copula of $\delta$ is selected, but this must be done cautiously and in a way that ensures $\delta$ has covariance $\Sigma_\delta$. Let us denote the copula as $V$. The joint distribution of $\delta$ will result when we couple together three marginals using $V$. That is, the CDF $F_\delta$ of $\delta$ is given by

$$F_\delta(a, b, c) = V(\Phi(a), G_1(b), G_2(c)),$$

where $G_1$ and $G_2$ are the CDFs of the scaled chi-square and $t$ distributions, respectively. As in the bivariate case, simulation from the above distribution involves first simulating $(U_1, U_2, U_3)$ from $V$, and then applying the quantile functions of the marginals to each coordinate of this vector. That is, the final simulated vector will be $\delta = (\Phi^{-1}(U_1), G_1^{-1}(U_2), G_2^{-1}(U_3))$. To construct the vine $V$ the researcher decides to couple the uniform marginals $U_1$ and $U_2$ with a Clayton copula, and $U_2$ and $U_3$ with a Joe copula. The dependence parameters of each of these bivariate copulas is numerically determined as described in the previous section, so that $\text{corr}(\Phi^{-1}(U_1), G_2^{-1}(U_2)) = \text{corr}(G_2^{-1}(U_2), G_3^{-1}(U_3)) = 0.4$. The hard part is now to couple $U_1$ with $U_3$ such that $\text{corr}(\delta_1, \delta_3) = \text{corr}(\Phi^{-1}(U_1), G_3^{-1}(U_3)) = 0.3$, and to achieve this we next introduce the concept of a vine.

Vines are convenient graphical tree structure models that can be used to build up high-dimensional distributions from conditional two-dimensional copulas. Vines therefore decompose the multivariate copula into a hierarchy of bivariate copulas. A vine on $p$ variables can be represented as a set of connected trees $V = \{T_1, \ldots, T_{p-1}\}$, where the edges of tree $j$ are the nodes of tree $j + 1$, $j = 1, \ldots, p - 2$ and are used to facilitate the picking out of various distributional characteristics, see Figure 4 for our current illustration with $p = 3$. The first tree has the variables as its nodes, and an edge between two variables means that these two variables are unconditionally coupled as in the previous section. In our case, we chose at the beginning to couple $U_1$ with $U_2$ and to couple $U_2$ with $U_3$. This corresponds to the tree at the bottom of Figure 4. The second tree has the edges of the first tree as its nodes. In our case the first tree has only two edges: $U_1, U_2$ and $U_2, U_3$. The second tree must therefore join $U_1, U_2$ and $U_2, U_3$. This tree has one single edge, which is denoted by $U_1, U_3|U_2$. That is, the second tree specifies the copula between $U_1$ and $U_3$, conditional on $U_2$. Note that we could have chosen a different tree at the first level, with edges, say, $U_1, U_2$ and $U_1, U_3$, which would yield a different distribution. Also the bivariate copulas chosen for coupling pairs of variables could have been chosen differently, yielding other types of vine distributions. However, we do not explore the flexibility of vines in the present paper. We see in Figure 4 that there are a total of three edges in the vine, and that the edges correspond to the pairwise correlations among the three variables. This holds also in higher-dimensional vines: There is an exact correspondence between the edges in the set of trees, and pairs of variables. So each off-diagonal element in the covariance matrix corresponds to a unique edge in the vine. The researcher's goal now is to define the distribution of $U_1$ and $U_3$. As suggested in Figure 4, this is done by specifying the distribution of $U_1$ and $U_3$, conditional on $U_2$. The researcher chooses a Frank copula for this distribution.

In terms of the joint density function $f$ of $\delta = (\delta_1, \delta_2, \delta_3)^\top$, its general form is:

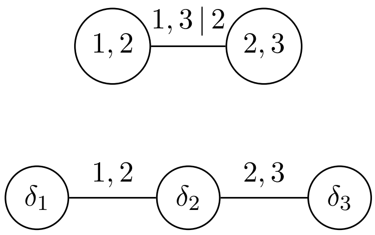


Figure 4: A three-dimensional regular vine.

$$f(a,b,c) = f_1(a)f_2(b)f_3(c)\cdot c_{12}(F_1(a),F_2(b))\cdot c_{13}(F_1(a),F_3(c))\cdot c_{13|2}(F_{1|2}(a|b),F_{3|2}(c|b)),$$

where $f_1, f_2, f_3$ are chosen marginal density distributions of $\delta_1, \delta_2, \delta_3$ respectively, where $c_{12}, c_{13}, c_{13|2}$ are the chosen bivariate copulas of $(\delta_1, \delta_2)$, $(\delta_1, \delta_3)$ and $(\delta_1, \delta_3)$ conditioned on $\delta_2$, respectively. Also, $F_1, F_2, F_3$ are CDFs of $\delta_1, \delta_2, \delta_3$ respectively, and $F_{1|2}, F_{3|2}$ are conditional CDFs of $\delta_1$ given $\delta_2$, and of $\delta_3$ given $\delta_2$, respectively. These CDFs are consequences of the chosen bivariate copulas and marginals. A full discussion with formulas for these conditional CDFs and of the joint density is included in the appendix. An advantage of vine distributions, compared to other multivariate simulation approaches where covariance matrices are specified (e.g, Ruscio and Kaczetow 2008; Qu *et al.* 2019), is the above explicit formula for the distribution of the simulated vector.

Returning to the illustrative example, we sum up the researcher's specifications for the residual error vector $\delta$:

- $\delta_1$ follows a standard normal distribution, $\delta_2$ follows a scaled chi-square distribution with one DF, and $\delta_3$ follows a scaled $t$ distribution with five DFs.
- The vine structure is given in Figure 4.
- A Clayton copula density for $c_{12}$, with dependence parameter calibrated so that $\Phi^{-1}(U_1)$ and $G_2^{-1}(U_2)$ have correlation 0.4.
- A Joe copula density for $c_{23}$, with dependence parameter calibrated so that $G_2^{-1}(U_2)$ and $G_3^{-1}(U_3)$ have correlation 0.4.
- A Frank copula density for $c_{13|2}$, with dependence parameter calibrated so that $\Phi^{-1}(U_1)$ and $G_2^{-1}(U_3)$ have correlation 0.3.

To construct multivariate distributions where the marginals and the covariances are prespecified, Grønneberg and Foldnes (2017) proposed the use of vines, resulting in the vine-to-anything (VITA) method. In our illustration, the researcher's requests may be fulfilled by constructing a VITA distribution using the **covsim** package as follows:

```
R> sigma.target <- matrix(c(1, 0.4, 0.3, 0.4, 1, 0.4, 0.3, 0.4, 1), 3)
R> margins <- list(list(distr = "norm"), list(distr = "chisq", df = 1),
+    list(distr = "t", df = 5))
R> pcs <- list(list(bicop_dist("clayton"), bicop_dist("joe")),
+    list(bicop_dist("frank")))
R> vine_cop <- vinecop_dist(pcs, structure = dvine_structure(1:3))
```

```
R> margin.variances <- c(1, 2, 5/3)
R> pre <- diag(sqrt(margin.variances/diag(sigma.target)))
R> vita.target <- pre %*% sigma.target %*% pre
R> set.seed(1)
R> calibrated.vita <- vita(margins, vita.target, vc = vine_cop,
+    verbose = TRUE)
```

```
Tree 1
    1 - 2 ( 1 of 3 )
    2 - 3 ( 2 of 3 )
Tree 2
    1 - 3 ( 3 of 3 )
```

```
R> post <- diag(1/diag(pre))
R> vita.sample <- rvine(10^5, calibrated.vita) %*% post
R> round(cov(vita.sample) - sigma.target, 2)
```

```
       [,1]   [,2]   [,3]
[1,] -0.001  0.001 -0.004
[2,]  0.001  0.001 -0.002
[3,] -0.004 -0.002  0.002
```

In the last lines of code above, we simulated a $n = 10^5$ sample from the calibrated VITA distribution using the function `rvine()` from the R package **rvinecopulib** (Nagler and Vatter 2021). The purpose of the last line is to confirm that the covariance matrix in the simulated sample is close to the target matrix.

A visualization of $n = 1000$ randomly drawn error vectors is presented in Figure 5. Note that, expectedly, the first marginal distribution is approximately standard normal, while the second and third marginal distributions are in accordance with scaled chi-square and Student's $t$ distributions.

Now, having constructed a VITA distribution for the residual errors, the researcher may use simulation to assess whether the quality of NTML inference for $\rho$, the correlation between the intercept and slope factors, deteriorates under non-normal residual errors. As a benchmark, the researcher simulates from a fully normal distribution on the observed variables $x_1$, $x_2$, and $x_3$. For the non-normal case, the researcher first simulates VITA residual errors, and then combines these with simulated intercept and slope values, each drawn from standard normal distributions, to obtain simulated observations on $x_1, x_2$ and $x_3$. The researcher replicates 1000 samples of size $n = 1000$ from both the fully normal distribution and the distribution with residual errors stemming from VITA.[1] The growth model was estimated with seven free parameters: three correlated residual errors; the residual error variances, which were constrained to be the same at each of the three measurement occasions; the correlation $\rho$; and the means of the intercept and slope variables. The model has one DF. As a measure of inference quality for $\rho$ the researcher decides to calculate the confidence interval coverage rate, at the 95% confidence level, for $\rho = 0$, using classical standard errors that assume exact

[1] Simulation code is provided in the online supplementary material.

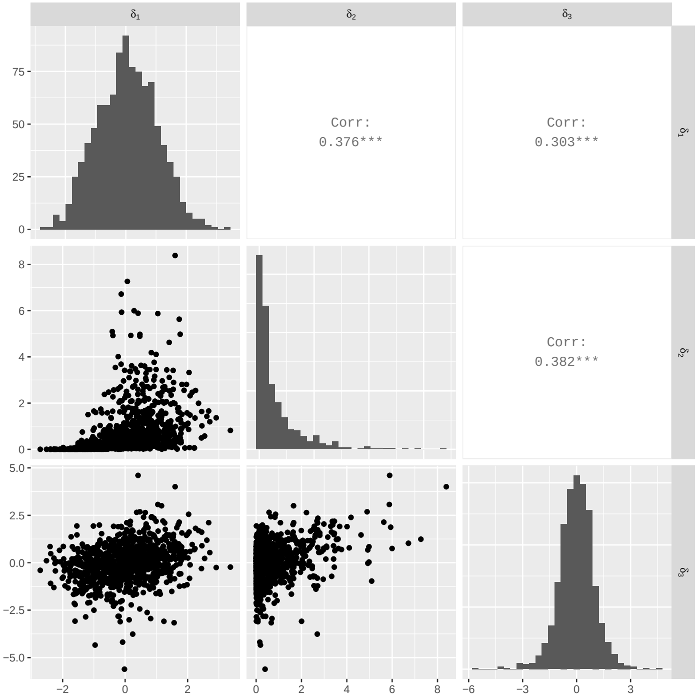


Figure 5: Scatterplots and histograms for a $n = 1000$ sample drawn from a three-dimensional VITA distribution.

normality. Under full normality, the coverage rate of 0.94 was close to nominal. With a non-normal error vector, the coverage rate was 0.905. Hence, the researcher found some support for the claim that non-normality in the error vector may affect the quality of intercept-slope correlation NTML inference.

## 3.1. The independent generator approach

The **covsim** package exports, in addition to `vita()`, the function `rIG()`. This simulation function is not based on a copula perspective and does not allow for full specification of the univariate marginal distributions. Instead it is closer in approach to the method of Vale and Maurelli (Vale and Maurelli 1983), where only univariate skewness and kurtosis are prespecified. However, the independent generator (IG) algorithm (Foldnes and Olsson 2016) is more flexible than the Vale and Maurelli method, defining a larger class of non-normal distributions for each set of skewness and kurtosis values. Although the main focus of the present manuscript is the flexible use of bivariate copulas in simulating non-normal data with given marginals and covariance matrix, we here for completeness give a short introduction to the IG algorithm.

The IG transform represents the non-normal vector $\xi$ stochastically as

$$\xi = AX,$$

where $A$ is a square matrix and $X$ a vector consisting of mutually independent generator variables with unit variance. The user specifies desired skewness and kurtosis values in $\xi$, and the IG algorithm numerically determines the skewness and kurtosis in each generator variable

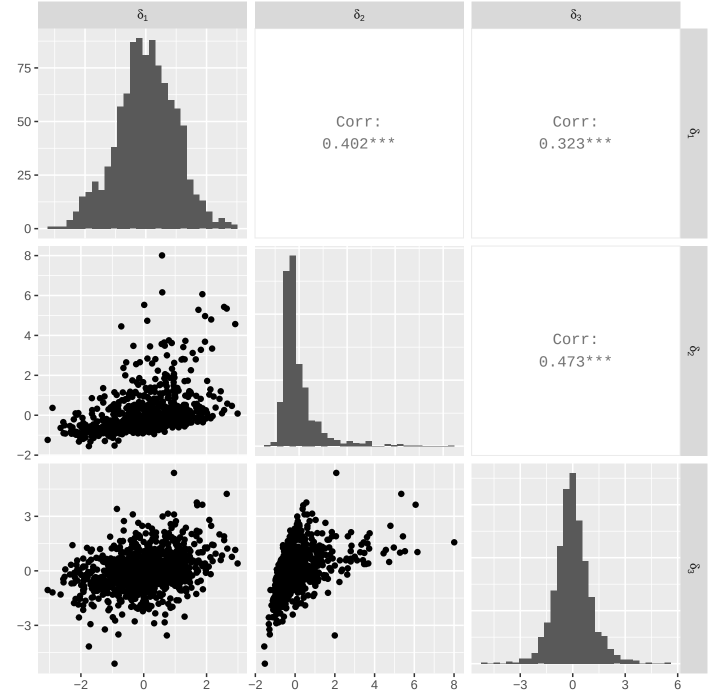


Figure 6: Scatterplots and histograms for a $n = 1000$ sample drawn from a three-dimensional IG distribution.

to match the desired values. The matrix $A$ is a square root of the specified covariance matrix $\Sigma$. In `rIG()` the user may specify a triangular square or a symmetrical square root matrix, which gives two different distributions. Also the marginal distributions for $X$ may be freely chosen, further expanding the distributional class defined by IG. In its current implementation, `rIG()` uses the Pearson family of distributions (Pearson 1895). Let us reconsider the marginal distributions used above. We note that the chi-square distribution with one DF and the Student's $t$ distribution with five DFs have skewness $\sqrt{8}$ and 0, respectively, and kurtosis 12 and 6, respectively. In the following code we ask for an IG distribution that matches the first four moments of the three marginal distributions considered in the previous section. The scatterplot is given in Figure 6.

```
R> set.seed(1)
R> ig.sample <- rIG(N = 10^3, sigma.target = sigma.target, reps = 1,
+    skewness = c(0, sqrt(8), 0), excesskurtosis = c(0, 12, 6))
```

# 4. A six-dimensional growth curve illustration

In this section we use the flexibility of VITA to further study the effect of non-normality in growth curve residual error vectors on normal-theory based inference. We focus on the chi-square statistic of model fit. We consider a linear growth curve with scores across six time-points. We assume that the errors $\delta_i$, $i = 1, \ldots, 6$, have unit variance, and that they are autocorrelated according to the following banded structure (i.e., a Toeplitz structure):

$$\Sigma_\delta = \begin{pmatrix} 1 & & & & & \\ 0.5 & 1 & & & & \\ 0.2 & 0.5 & 1 & & & \\ 0 & 0.2 & 0.5 & 1 & & \\ 0 & 0 & 0.2 & 0.5 & 1 & \\ 0 & 0 & 0 & 0.2 & 0.5 & 1 \end{pmatrix}$$

We calibrate the following three VITA distributions for the $\delta$ vector:

**VITA1** $\delta_1, \delta_2, \delta_3, \delta_4, \delta_5$, and $\delta_6$ are standard normal, and all 18 bivariate copulas are of Clayton type.

**VITA2** $\delta_1$ is standard normal, while $\delta_2, \delta_3, \delta_4, \delta_5$, and $\delta_6$ are chi-square distributed with $5, 4, 3, 2$, and 1 degrees of freedom, respectively. The chi-square distributions are scaled to have unit variance. All 18 bivariate copulas are normal.

**VITA3** $\delta_1$ is standard normal, while $\delta_2, \delta_3, \delta_4, \delta_5$, and $\delta_6$ are chi-square distributed with $5, 4, 3, 2$, and 1 degrees of freedom, respectively. The chi-square distributions are scaled to have unit variance. All 18 bivariate copulas are of type Clayton.

Using the `vita()` function in package **covsim** the code is as follows:

```
R> residual.covariance <- toeplitz(1:6)
R> residual.covariance[residual.covariance > 3] <- 0
R> residual.covariance[residual.covariance == 2] <- 0.5
R> residual.covariance[residual.covariance == 3] <- 0.2
R> margins.nonnorm <- list(list(distr = "norm"),
+    list(distr = "chisq", df = 5), list(distr = "chisq", df = 4),
+    list(distr = "chisq", df = 3), list(distr = "chisq", df = 2),
+    list(distr = "chisq", df = 1))
R> margins.norm <- list(list(distr = "norm"), list(distr = "norm"),
+    list(distr = "norm"), list(distr = "norm"),
+    list(distr = "norm"), list(distr = "norm"))
R> margin.variances <- c(1, 10, 8, 6, 4, 2)
R> sigma.target <- diag(sqrt(margin.variances)) %*% residual.covariance %*%
+    diag(sqrt(margin.variances))
R> set.seed(1)
R> vita1 <- vita(margins.norm, residual.covariance, family_set = "clayton")

Tree 1
    1 - 2 ( 1 of 15 )
[...]

R> set.seed(1)
R> vita2 <- vita(margins.nonnorm, sigma.target, family_set = "gauss")
```

```
Tree 1
    1 - 2 ( 1 of 15 )
[...]

R> set.seed(1)
R> vita3 <- vita(margins.nonnorm, sigma.target, family_set = "clayton")

Tree 1
    1 - 2 ( 1 of 15 )
[...]
```

Data generation first simulates independent random draws from the standard normal for the intercept and slope variables, and then adds the residual errors simulated from VITA distributions. A growth model with 15 degrees of freedom, which correctly specifies the structure for $\Sigma_\delta$, is fitted to the data. Our research question is to what extent non-normality in the errors affects the sampling distribution, and in particular, the type I error control of the regular normal-theory chi-square statistic $T_{\text{NTML}}$. Since VITA1, VITA2, and VITA3 are different distributions, with different mixes of marginal and copula non-normality, there might also be insights to draw from their differential effect on the chi-square test. One way of conducting this research is to use conventional small-sample simulations and to calculate rejection rates over many replications. Here we choose a different approach. We calculate the exact asymptotic distribution of $T_{\text{NTML}}$ in each distributional condition. This will also give us asymptotic type I error rates.[2] First, we simulate a very large $n = 10^6$ sample from each of the VITA distributions. Then the model is fitted to each of the three datasets, and we extract the eigenvalues of the matrix $U\Gamma$ (see, e.g., Foldnes and Grønneberg 2018 for further details). Theory (Box 1954) dictates that $T_{\text{NTML}}$ is asymptotically distributed as the weighted sum of independent chi-square distributions, each with one degree of freedom, where the weights are the eigenvalues of $U\Gamma$. This allows us to calculate the density of $T_{\text{NTML}}$ under the three distributional conditions. Under multivariate normality, this density is that of the nominal chi-square distribution with 15 degrees of freedom, which is used to calculate asymptotic type I error control of $T_{\text{NTML}}$. Figure 7 depicts the asymptotic sampling distribution of $T_{\text{NTML}}$ under four conditions, namely multivariate normality, and the distributions involving the three VITA error distributions. It is seen that $T_{\text{NTML}}$ becomes inflated as we move from multivariate normality, and as we progress through the three VITA error distributions. The vertical line represents the critical value when referring $T_{\text{NTML}}$ to its critical value at the $\alpha = 0.05$ level of significance. VITA1 has standard normal marginals and a non-normal copula. In this condition the asymptotic type I error control is 7.3%, quite close to the nominal level. VITA2 has four non-normal marginals, and a normal copula, and affects $T_{\text{NTML}}$ to a larger extent than VITA1. The asymptotic rejection rate under VITA2 is 29.4%, which is far above the nominal 5% level. VITA3 introduces more non-normality compared to VITA2, by having a non-normal copula. The effect on $T_{\text{NTML}}$ is critical, whose asymptotic rejection rate is 50.6% under VITA3 errors. In sum we see that non-normality in the residual error vector may markedly inflate the rejection rates of $T_{\text{NTML}}$, but we may speculate that the effect is mild as long as the univariate marginals are normally distributed.

[2] In the online supplementary material are given code for conventional small-sample simulations that confirms our upcoming findings.

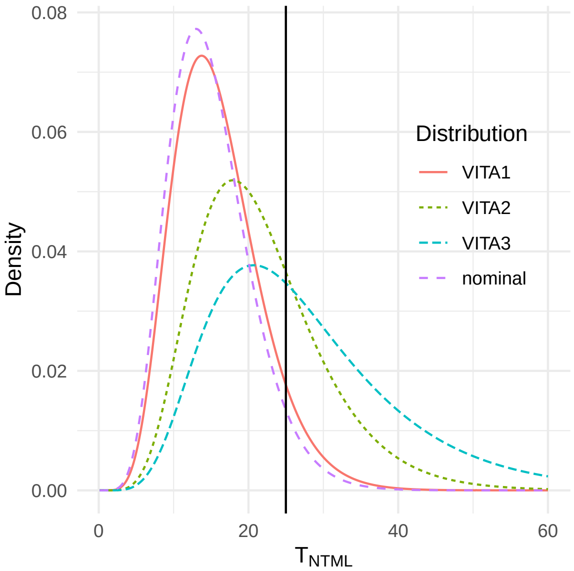


Figure 7: The asymptotic density of $T_{\text{NTML}}$ under four conditions. nominal = Chi-square distribution with 15 degrees of freedom. VITA1, VITA2, and VITA3 denote three kinds of non-normal residual distributions. Vertical line represents critical value at $\alpha = 0.05$.

## 5. The implementation of VITA in covsim

In this section we briefly explain how the function `vita()` implements the VITA algorithm. Grønneberg and Foldnes (2017) provided as supplementary material a VITA implementation using package **VineCopula** (Schepsmeier, Stoeber, Brechmann, Graeler, Nagler, and Erhardt 2021) for constructing and simulating from regular vines. This package is no longer in active development, and package **rvinecopulib** (Nagler and Vatter 2021) was instead used in `vita()`. The most important benefits of **rvinecopulib** relative to **VineCopula** for our purposes is a sleeker and more modern application programming interface (API) and shorter simulation runtimes. In experiments (see supplementary material) with a five-dimensional vine, on a computer with 4 CPU cores, simulation runtimes at a sample size of $n = 1000$ were shorter with **rvinecopulib** compared to **VineCopula** by a factor of four. Also, as explained below, the initial calibration of VITA parameters involves a series of large-sample random draws from regular vines, which means that VITA calibration is computationally demanding. The root-finding routine provided by Grønneberg and Foldnes (2017) has been improved in `vita()`, by splitting it into a high-speed routine which identifies an interval for the root, followed by high-precision root-finding in this interval, based on curve-fitting. This two-stage root-finding routine is faster than the basic method in Grønneberg and Foldnes (2017). Combined with faster simulation times in package **rvinecopulib**, the calibration time for a five-dimensional vine using `vita()` instead of the original code provided by Grønneberg and Foldnes (2017) was reduced by a factor of 13.

The main arguments to `vita()` are

- `margins`. A list that specifies the univariate marginal distributions.
- `sigma.target`. The target covariance matrix.

- `vc`. A vine copula structure in the format defined by package **rvinecopulib**. That is, a specification of a hierarchy of $p-1$ trees, and, for each tree node, a bivariate copula family. If not provided by the user, `vita()` will initialize `vc` as follows. The vine structure of `vc` is specified as the simplest regular vine, namely the D-vine on $p$ dimensions. See Figures 4 and 10 for the D-vine with $p = 3$ and $p = 4$, respectively. In addition, the bivariate copula family in each node in the D-vine will be taken as the first element of the argument `family_set`.

- `family_set`. A vector that specifies which bivariate copula families are to be calibrated. If `vc` is provided by the user, and the algorithm can not identify a feasible solution for the family dictated by `vc`, the algorithm instead tries to calibrate the dependence parameter for the first family in `family_set`. If not successful, an attempt is made to calibrate the parameter in the second family, and so forth. If `vc` is not provided, the algorithm attempts first to calibrate the dependence parameter in the first member of `family_set`, and if not successful, the second member, and so forth.

The above arguments specify a class of VITA distributions, parameterized by the $p(p-1)/2$ dependence parameters in the bivariate copulas. The task of `vita()` is to numerically determine the values of these dependence parameters so that the resulting VITA distribution has the required covariance matrix given by `sigma.target`. That is, `vita()` searches for a dependence parameter value $\theta$ in each of the copulas, so that the covariance matrix of the resulting full distribution is `sigma.target`, up to numerical precision. This search can be done in the same order as when simulating from a vine, building our way up the tree, connecting more and more distributions with pairwise conditional distributions. As shown in Grønneberg and Foldnes (2017), the correlation of each pair of variables is typically a strictly increasing function of $\theta$ for most single parameter copulas, making the numerical search well behaved. Unfortunately, there is no simple formula for the correlation matrix of a vine. Worse still, no simple formula can be derived for pairwise bivariate distributions connected at higher levels of the vine tree. In the implementation of VITA in `vita()`, we resort to Monte Carlo simulation to approximate the required correlation.

VITA calibrates each pairwise bivariate distribution and combines them to form the full vine distribution in a specific order. A formal algorithmic description of the methodology is given in Grønneberg and Foldnes (2017). We here informally summarize the main steps of the method, and later describe in technical detail the new implementation of the root finding procedure that underlies the calibration.

As explained in more detail in the appendix, each pair $(i, j)$, where $1 \leq i, j \leq p$, is connected once in the vine. This is also the case in the covariance matrix. Let $(\sigma_{ij})$ be the target covariance matrix in `sigma.target`, and let the parameter of the bivariate copula connecting the $(i, j)$ distribution be parameterized by $\theta_{ij}$. The first pairs of bivariate distributions that are calibrated are those connected at the lowest level of the vine tree. For illustration, we consider the vine given in Figure 4 (p. 10). We see that $(1, 2)$ are connected at the lowest tree. We may therefore simulate directly from this bivariate distribution. This distribution depends on a parameter $\theta_{12}$. We may choose this parameter in such a way that the covariance of the resulting bivariate distribution matches the required covariance given in $\sigma_{12}$. This matching is non-trivial and is described in technical detail below. A similar matching may be done for all other marginals connected at the lowest level of the vine, which here is only $(2, 3)$. These calibrations are done independently of each other. Now the vine in Figure 4 has only two levels, and there is only one bivariate margin left to be matched, namely $(1, 3)$, which is connected at the topmost level. The distribution of $(1, 3)$ is derivable from the full vine structure, and the conditional copula of $(1, 3)$ given 2 is parameterized using a bivariate copula with parameter $\theta_{13}$. To simulate realizations from $(1, 3)$ we need to know the distributions at the lower level of the tree, as well as specifying the value of $\theta_{13}$. The distributions of the lowest level of the tree have already been fixed, and we may, for varying values of $\theta_{13}$, simulate the full three-dimensional vine distribution, compute the covariance of $(1, 3)$, and select the $\theta_{13}$ value that yields a covariance equal to $\sigma_{13}$. We have then calibrated the full three-dimensional vine.

For higher dimensions, this idea has to be iterated several times to identify all parameters in an order that enable us to always simulate from the required bivariate variables in the vine. In order to briefly illustrate how to calibrate a higher-dimensional vine, consider the four-dimensional vine in Figure 10 in the appendix (p. 41). Comparing the vine in Figures 4 and 10, we see that the three-dimensional vine in Figure 4 is included within the structure of the vine of Figure 10 as a subset of its connections. That is, the vine in Figure 4 is a sub-vine of the vine of Figure 10 that comprises the variables $(1, 2, 3)$: The vines are equal, with the exception that the four-dimensional vine also has to connect marginal 4 with the remaining variables, which is done using the additional structure given in Figure 10. But to simulate only $(1, 2, 3)$ from the four-dimensional vine in Figure 10, we only need to know the three-dimensional vine in Figure 4.

To calibrate the four-dimensional vine in Figure 10, we may therefore continue where we left off when calibrating the three-dimensional vine in Figure 4. After calibrating the new bivariate distribution connected at the lowest level, namely $(3, 4)$, the next step is to calibrate all distributions at the second level. Only one such distribution is left, namely $(2, 4)$. By the same reasoning as earlier, we may simulate from the sub-vine that enables the simulation of $(2, 3, 4)$: The parameters of the $(2, 3)$ and $(3, 4)$ distributions have already been fixed. Therefore, the sub-vine expressing the full distribution of $(2, 3, 4)$ only have one free parameter, namely $\theta_{24}$, which may be varied to get a covariance between $(2, 4)$ to match up with $\sigma_{24}$. As in the first level of the tree, the matching of the parameters at the second level of the tree are done independently of each other, and can be done in any order. However, to calibrate all connections at a given level, all connections at lower levels have already to be calibrated from before.

The final matching required for the four-dimensional vine is then to work with the single distribution connected at the third level of the vine, namely the $(1, 4)$ distribution, so that the $(1, 4)$ marginal has covariance equal to $\sigma_{14}$. All parameters of the vine are now fixed with the exception of $\theta_{14}$, and this parameter may be varied until the required covariance is induced.

The calibration order of variable pairs in `vita()` is as follows: All copulas at the lowest level are calibrated, then the next level, and so on up to the highest level. As mentioned in Grønneberg and Foldnes (2017), other orders are possible, as exemplified above, but the order is immaterial as long as unique solutions for reaching the desired covariances exist.

In each calibration step, numerical integration done via Monte Carlo simulation and a search for the solution of an equation must be performed. We now detail how this is done in the implementation of `vita()`. Let $(U_i, U_j)$ be distributed according to the copula of the sub-vine required to simulate the $(i, j)$ distribution as described above. Due to the order we have traversed the vine, there is always only one free parameter $\theta_{ij}$ of this distribution that is free, and is used to match up the required covariance $\sigma_{ij}$. In the following description we omit the $ij$ subscript from $\theta$, to reduce the notational burden. As explained in more detail in the appendix, we apply the corresponding inverse quantile functions according to the entries in `margins` to calculate the covariance induced by a given $\theta$: Let us denote by $\sigma_{ij}(\theta)$ the covariance between the resulting variables $F_i^{-1}(U_i)$ and $F_j^{-1}(U_j)$. Our aim is now to determine $\theta$ so that $\sigma_{ij}(\theta) = \sigma_{ij}$. Unfortunately there are no analytical expressions available for $\sigma_{ij}(\theta)$ except in very special cases, but the covariance may be approximated by simulating a large sample of size $n$ of $(U_i, U_j)$, applying the quantile functions $F_i^{-1}$ and $F_j^{-1}$ to each simulated variable, and calculating the resulting sample covariance, which we denote by $\hat{\sigma}_{ij}(\theta)$. Then $\hat{\sigma}_{ij}(\theta)$ will converge in probability to $\sigma_{ij}(\theta)$ as $n$ increases. However, with large $n$, simulating these samples is time-consuming, so `vita()` is implemented in two stages.

1. *Initial high-speed calibration.* In this stage we use the modest sample size $n = 1500$ to determine $\hat{\sigma}_{ij}(\theta)$, using function `uniroot()` in the **stats** package. That is, we approximate $\theta$ by finding the root $\hat{\theta}_n$ of the discrepancy function $\hat{\sigma}_{ij}(\theta) - \sigma_{ij}$. We expect $\hat{\theta}_n$ to be quite close to $\theta$, but it contains random error, so we repeat this procedure and approximate $\theta$ a small number of times (the argument `numrootpoints`). This results in a number of root candidates $\hat{\theta}_n^1, \hat{\theta}_n^2, \ldots, \hat{\theta}_n^{\texttt{numrootpoints}}$, which are independent and identically distributed random variables. A standard $t$-based confidence interval for the dependence parameter $\theta$ is then constructed from these approximate roots, using a high level of confidence (as specified by the argument `conflevel`).

2. *Final high-precision calibration.* In this stage, we evaluate the discrepancy $\hat{\sigma}_{ij}(\theta) - \sigma_{ij}$ to a high precision at a small number (as specified by argument `numpoints`) of equally spaced points across the confidence interval determined in the first stage. The approximation is done by simulating from a very large sample ($n$ is equal to the argument `Nmax`) in each of these points. We then fit a second degree polynomial to the discrepancy values and use `uniroot()` to locate the root of this polynomial, which yields our final estimate for $\theta$.

If, for any pair of variables, the calibration does not find a solution $\theta$, the algorithm changes the bivariate family to the next entry in `family_set`. If no solution is found, `vita()` terminates with an error message. This means that there is no VITA distribution with the given marginals, vine structure and bivariate families that can attain `sigma.target`. To proceed, the user could then rerun `vita()` with, e.g., a different vine structure.

As mentioned in the introduction, traditional approaches to non-normal covariance modeling only specify the lower-order univariate moments, and do not offer any control of the multivariate aspects of the simulated vector, with the exception of covariance matching. As we have demonstrated, the VITA approach is more flexible. However, the cost of increased flexibility is increased computing time necessary to calibrate the VITA distribution. The default values for arguments `Nmax` and `numpoints` in `vita()` guarantee a highly precise VITA calibration. That is, the calibrated VITA distribution will have a covariance matrix almost numerically indistinguishable from `sigma.target`. In higher dimensions, this precision comes at the cost of long calibration running times. Table 1 gives calibration times on a computer (2.3 GHz 8-Core Intel Core i9) using the default options in `vita()`, with target correlation among all variable pairs equal to $\rho = 0.3$, for increasing dimensionality. It is seen that approaching 20

| Dimension | 5 | 10 | 15 | 20 | 25 | 30 |
|---|---|---|---|---|---|---|
| Calibration (hrs) | 0.006 | 0.065 | 0.401 | 1.447 | 3.675 | 8.100 |
| Simulation (hrs) | 0.001 | 0.004 | 0.009 | 0.015 | 0.024 | 0.034 |

Table 1: Calibration times in hours under the default `Nmax` $= 10^6$. Simulation of 1000 samples, each of size $n = 10^3$.

dimensions, calibration time exceeds one hour, while simulating 1000 samples, each of size $n = 1000$, requires less than a minute. So the calibration step, which is only executed once, is time-consuming, while repeated simulation from the calibrated VITA is relatively fast. Foldnes and Grønneberg (2022b) calibrated and simulated from VITA distributions in twenty dimensions in an extensive simulation design. However, given that the median number of observed variables in empirical SEM studies is close to 20 (Li 2016), using `vita()` for larger models, with say 50 dimensions, will entail days of calibration time with the default options. In such cases the user may lower the argument `Nmax` from $10^6$ to $10^5$, thereby reducing calibration time by a factor of 10. For instance, for dimension 40 calibration was achieved in 6.2 hours using option `Nmax` $= 10^5$. Even with reduced precision, the calibrated VITA distribution has a covariance matrix almost equal to the target covariance. Among the 780 pair-wise correlations estimated in a $n = 10^6$ sample drawn from the 40-dimensional calibrated VITA distribution with `Nmax` $= 10^5$, 737 were within a 0.005 distance of the target $\rho = 0.3$, and all (except for one) were within a 0.01 distance of $\rho = 0.3$.

If high precision is important in an application, a formal test of equality of covariance matrices should be performed. This may be done by computing as test statistic the quadratic form of the discrepancies between the sample covariances and the target covariances, weighted by the inverse of the estimated asymptotic covariance matrix of the covariances (Mair *et al.* 2012).

To precisely (`Nmax` $= 10^6$) calibrate VITA distributions with 50 or more dimensions, our current implementation will demand unrealistically long running times. The bottleneck of the calibration algorithm consists of simulating a large sample (`Nmax`) from a regular vine. This simulated sample is then used to compute a single covariance. If we distribute the large-sample simulation to several computers, the desired covariance of all the simulated realizations across computers can be computed based on sums, cross-products and sums of squares from each computer. Hence, the VITA algorithm may conveniently be distributed across a network of computers. Such functionality may be included in future versions of **covsim**.

# 6. Further examples

We here consider some further applications of VITA. In Section 6.1, we consider a 20-dimensional SEM example with continuous data. In Section 6.2 we discuss simulation of ordinal SEMs and show how this can be done with VITA.

## 6.1. Using VITA to simulate continuous data for SEM

In most SEM simulation studies the methodologist first specifies a SEM model together with its population parameter values. Then the study is conducted by drawing random samples from a distribution whose covariance matrix equals the model-implied covariance matrix. As an example, consider the SEM whose structural part is depicted in Figure 8. The model

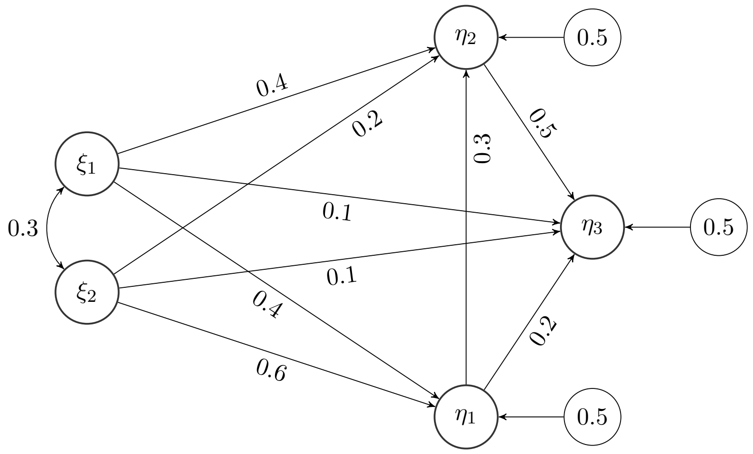


Figure 8: Structural model for a medium-sized SEM. Indicator variables not depicted.

has five factors, and is representative of a medium-sized SEM in applied studies (Li 2016). Each factor has four indicators, yielding a total of 20 dimensions. The factor loadings for the indicators were set to $0.8, 0.7, 0.6$ and $0.5$ within each factor, and the corresponding residual variances were set to 0.5. The correlation was set to 0.3 between the two exogenous factors, each of which had unit variance. The residual variances for the endogenous factors were also set to 0.5. Using the package **lavaan** we can compute the target covariance matrix implied by these population parameters as follows.

```
R> sem.pop <- '
+    Ksi1 =~ start(.8) * x1 + start(.7) * x2 + start(.6) * x3 +
+      start(.5) * x4
+    Ksi2 =~ start(.8) * x5 + start(.7) * x6 + start(.6) * x7 +
+      start(.5) * x8
+    Eta1 =~ start(.8) * y1 + start(.7) * y2 + start(.6) * y3 +
+      start(.5) * y4
+    Eta2 =~ start(.8) * y5 + start(.7) * y6 + start(.6) * y7 +
+      start(.5) * y8
+    Eta3 =~ start(.8) * y9 + start(.7) * y10 + start(.6) * y11 +
+      start(.5) * y12
+    Eta1 ~ start(.4) * Ksi1 + start(.6) * Ksi2
+    Eta2 ~ start(.4) * Ksi1 + start(.2) * Ksi2 + start(.3) * Eta1
+    Eta3 ~ start(.1) * Ksi1 + start(.1) * Ksi2 + start(.2) * Eta1 +
+      start(.5) * Eta2
+    Ksi1 ~~ start(.3) * Ksi2; Eta1 ~~ start(.5) * Eta1;
+    Eta2 ~~ start(.5) * Eta2; Eta3 ~~ start(.5) * Eta3
+    x1 ~~ start(.5) * x1; x2 ~~ start(.5) * x2
+    x3 ~~ start(.5) * x3; x4 ~~ start(.5) * x4; x5 ~~ start(.5) * x5
+    x6 ~~ start(.5) * x6; x7 ~~ start(.5) * x7; x8 ~~ start(.5) * x8
+    y1 ~~ start(.5) * y1; y2 ~~ start(.5) * y2; y3 ~~ start(.5) * y3
+    y4 ~~ start(.5) * y4; y5 ~~ start(.5) * y5; y6 ~~ start(.5) * y6
+    y7 ~~ start(.5) * y7; y8 ~~ start(.5) * y8; y9 ~~ start(.5) * y9
+    y10 ~~ start(.5) * y10; y11 ~~ start(.5) * y11; y12 ~~ start(.5) * y12'
```

```
R> sigma.target <- lavInspect(sem(sem.pop, data = NULL), "sigma.hat")
```

Next, we fit a VITA distribution with normal marginals to the target covariance matrix. This is a variant of a data generating distribution used in the simulation study of Foldnes and Grønneberg (2022b). First, the margins are scaled to match the target variances. Then, we calibrate a VITA distribution. Note that we do not specify which family of copulae to use, so the default Clayton copula is used. Finally, a list of 1000 samples, each of sample size 1000, is drawn from the calibrated vita distribution.

```
R> marginsnorm <- lapply(X = sqrt(diag(sigma.target)),
+    function(X) list(distr = "norm", sd = sqrt(X)))
R> vitadist <- vita(marginsnorm, sigma.target)

Tree 1
    1 - 2 ( 1 of 190 )
[...]

R> randomsamples <- replicate(10^3, rvine(10^3, vitadist))
```

As discussed previously, the calibration step is time-consuming in higher dimensions. Here, with 20 variables, the calibration step required 1.8 hours (again using a 2.3 GHz 8-Core Intel Core i9 CPU). This step is only performed once. When completed, random samples can be drawn at a relatively fast rate. Producing 1000 samples each of size 1000 took one minute to complete. Finally, we note that the calibration step may be performed faster by specifying option `Nmax` $= 10^5$ when calling `vita()`, at the expense of reduced precision in covariance matching.

### 6.2. Using VITA to simulate ordinal-categorical data for SEM

A major approach for SEM with ordinal data is to impose a threshold model to the data, which postulates that the categorical data arise from discretization of an underlying continuous vector which is multivariate normally distributed. Many influential simulation studies (e.g., Quiroga 1994; Rhemtulla, Brosseau-Liard, and Savalei 2012; Flora and Curran 2004; Li 2016) have investigated the robustness of ordinal SEM against violation of non-normality, using the Vale-Maurelli approach. However, Grønneberg and Foldnes (2019a) showed that the Vale Maurelli approach is not suitable for ordinal data simulation in the context of covariance modeling. We here briefly show how to simulate ordinal bivariate data by discretizing bivariate VITA distributions. For an observed ordinal variable, there is no way to identify which underlying univariate distribution produced the data, since the thresholds may be transformed to accommodate all continuous univariate distributions.

As argued in Foldnes and Grønneberg (2019a, 2022b), it is advantageous to keep the marginals fixed during simulation. Since VITA offers exact control of marginals, it is uniquely suited for simulation studies with ordinal data for SEM. We will here set the marginals to standard normal. When simulating fully normal data, both the marginals and the copula are normal. We will let the copula be non-normal, but the marginals will be normal.

For illustration, we assume that the underlying correlation in a continuous bivariate distribution with standard normal marginals is $\rho = 0.5$, and we discretize into three categories using

|   | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0.334 | 0.123 | 0.043 |
| 2 | 0.123 | 0.146 | 0.072 |
| 3 | 0.043 | 0.072 | 0.044 |

|   | 1 | 2 | 3 |
|---|---|---|---|
| 1 | 0.333 | 0.146 | 0.021 |
| 2 | 0.146 | 0.149 | 0.047 |
| 3 | 0.021 | 0.047 | 0.091 |

Table 2: Population contingency tables after discretizing the distribution with standard normal marginals and a Clayton copula (left) and the distribution with standard normal marginals and a Joe copula (right).

thresholds $\tau_1 = 0$ and $\tau_2 = 1$. This means that we consider simulated data of the form

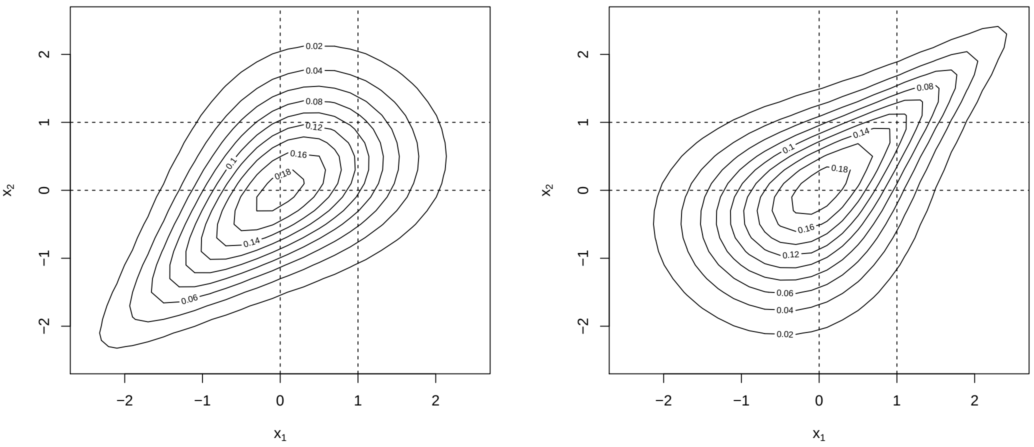
$$X_i = \begin{cases} 1, & \text{if } \xi_i \leq \tau_1 \\ 2, & \text{if } \tau_1 < \xi_i \leq \tau_2 \\ 3, & \text{if } \xi_i > \tau_2 \end{cases} = \begin{cases} 1, & \text{if } \xi_i \leq 0 \\ 2, & \text{if } 0 < \xi_i \leq 1 \\ 3, & \text{if } \xi_i > 1 \end{cases}$$

for $i = 1, 2$, where $(\xi_1, \xi_2)$ is a continuous random vector simulated using VITA. Both ordinal variables have proportions $0.5, 0.34$, and $0.16$. We inquire whether the polychoric correlation estimator used in ordinal SEM becomes biased when we replace the bivariate normal with a Clayton or a Joe copula. So first, we determine parameters for the latter two copulas such that, when marginals are standard normal, the Pearson correlation is 0.5.

```
R> sigma.target <- matrix(c(1, 0.5, 0.5, 1), 2)
R> set.seed(1)
R> vita_clayton <- vita(list(list(distr = "norm"), list(distr = "norm")),
+    sigma.target, family_set = "clayton")

Tree 1
    1 - 2 ( 1 of 1 )

R> set.seed(1)
R> vita_joe <- vita(list(list(distr = "norm"), list(distr = "norm")),
+    sigma.target, family_set = "joe")

Tree 1
    1 - 2 ( 1 of 1 )

R> clayton.disc <- apply(rvine(10^3, vita_clayton), 2, cut,
+    breaks = c(-Inf, 0, 1, Inf), labels = FALSE)
```

Contour plots of these calibrated copulas are given in Figure 9.

After discretizing the distribution with standard normal marginals and a Clayton copula shown in Figure 9a and the distribution with standard normal marginals and a Joe copula in Figure 9b, the resulting population contingency tables are given in Table 2. The two contingency tables in Table 2 indicate that under the Clayton and Joe copula the probabilities that both ordinal variables take their maximum value are 4.4% and 9.1%, respectively. Such discrepancies in the bivariate ordinal distribution affects the normal-theory based polychoric



(a) Clayton copula.

(b) Joe copula.

Figure 9: VITA distributions with standard normal marginals. The dashed lines represent thresholds $\tau_1 = 0$ and $\tau_2 = 1$ used for discretization.

estimator: The population value of the polychoric correlation under the Clayton and Joe copula is 0.42 and 0.60, respectively. Given that the true underlying Pearson correlation for both distributions in Figure 9 is 0.5, this shows that the polychoric correlation may become strongly (downwards or upwards) biased when underlying normality is violated. In this illustration, the lower tail dependency in the Clayton copula, combined with the chosen thresholds, results in strong downward bias. And the upper tail dependency in the Joe copula leads to strong upward bias.

The sensitivity of the polychoric correlation to underlying non-normality poses a threat to the popular practice of conducting SEM with ordinal data based on the polychoric correlation matrix. Even though a proposed SEM model fits well to the underlying data, it might fit poorly to the polychoric correlation matrix, since the latter might be biased due to non-normality in the underlying data. The result might be biased estimates and inflated type I error rates, and it is therefore important to assess whether the ordinal dataset is compatible with the underlying normality assumption. Foldnes and Grønneberg (2019b) proposed a bootstrap test for this purpose, which is implemented in the R package **discnorm** (Foldnes and Grønneberg 2021). It is used as follows.

```
R> library("discnorm")
R> bootTest(clayton.disc)
```

We conducted a small simulation study on the type I error control and power of the bootstrap test in the context of the present bivariate illustration. We generated 500 samples of size $n = 100$ and of size $n = 500$ and collected the rejection rate of the bootstrap test for ordinal data stemming from discretization of a normal distribution, the Clayton VITA, and the Joe VITA, using the same set of thresholds as depicted in Figure 9. The rejection rates are given in Table 3. The bootstrap test maintains type I error rates well, but has low power to detect the underlying normality in both the Clayton and Joe VITA distributions at the smallest sample size. Expectedly, at the larger sample size $n = 500$ the power to detect the underlying non-normality is higher. The non-normality in the Joe VITA is more detectable than the non-normality in the Clayton VITA at both sample sizes.

| | Underlying distribution | | |
|---|---|---|---|
| | Normal | Clayton | Joe |
| $n = 100$ | 0.088 | 0.162 | 0.242 |
| $n = 500$ | 0.052 | 0.584 | 0.852 |

Table 3: Rejection rates of the bootstrap test for underlying normality.

# 7. Conclusion

The VITA approach, implemented in the R package **covsim**, is very flexible, since it accommodates user-specified marginal distributions and also offers a great deal of control over the bivariate dependencies in the simulated vector. This control is in contrast to more established methodology, like the Vale and Maurelli (1983) method. In most cases, Vale-Maurelli has an exactly normal copula (Foldnes and Grønneberg 2015), and does not allow the specification of the resulting distribution except the covariance matrix, skewness and kurtosis. The increased flexibility of VITA, however, comes at a cost. VITA, being based on the statistical concept of a regular vine, is a more complex construction than the Vale Maurelli transform. Also, VITA calibration is more computationally demanding than is the case for the Vale-Maurelli transform. In the appendix we have given an introduction to the statistical machinery underlying vines and VITA, such as bivariate copulas and their use in constructing regular vines. We also give an introduction to how VITA simulation is performed on a computer. Numerical illustrations of applying VITA to simulate non-normal residual error structures in growth curve modeling were presented, demonstrating the effects of different kinds of non-normality on inference. Also, we have illustrated how VITA may be used to simulate continuous non-normal data from a SEM, and to simulate ordinal data in a way that properly violates the underlying normality assumption. For ordinal data, we illustrated a new bootstrap test, implemented in the R package **discnorm**, for the central assumption of underlying normality.

# References

Aas K, Czado C, Frigessi A, Bakken H (2009). "Pair-Copula Constructions of Multiple Dependence." *Insurance: Mathematics and Economics*, **44**(2), 182–198. doi:10.1016/j.insmatheco.2007.02.001.

Bedford T, Cooke RM (2002). "Vines – A New Graphical Model for Dependent Random Variables." *The Annals of Statistics*, **30**(4), 1031–1068. doi:10.1214/aos/1031689016.

Bentler PM (2006). "**EQS** 6 Structural Equations Program Manual."

Boomsma A (2013). "Reporting Monte Carlo Studies in Structural Equation Modeling." *Structural Equation Modeling: A Multidisciplinary Journal*, **20**(3), 518–540. doi:10.1080/10705511.2013.797839.

Box GEP (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effect of Inequality of Variance in the One-Way Classification." *The Annals of Mathematical Statistics*, **25**(2), 290–302. doi:10.1214/aoms/1177728786.

Cain MK, Zhang Z, Yuan KH (2017). "Univariate and Multivariate Skewness and Kurtosis for Measuring Nonnormality: Prevalence, Influence and Estimation." *Behavior Research Methods*, **49**(5), 1716–1735. doi:10.3758/s13428-016-0814-1.

Cario MC, Nelson BL (1997). "Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix." *Technical report*, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois.

Christoffersson A (1977). "Two-Step Weighted Least Squares Factor Analysis of Dichotomized Variables." *Psychometrika*, **42**(3), 433–438. doi:10.1007/bf02293660.

Curran PJ, West SG, Finch JF (1996). "The Robustness of Test Statistics to Nonnormality and Specification Error in Confirmatory Factor Analysis." *Psychological Methods*, **1**(1), 16–29. doi:10.1037/1082-989x.1.1.16.

Dimitrov DM (2002). "Reliability: Arguments for Multiple Perspectives and Potential Problems with Generalization across Studies." *Educational and Psychological Measurement*, **62**(5), 783–801. doi:10.1177/001316402236878.

Dissmann J, Brechmann EC, Czado C, Kurowicka D (2013). "Selecting and Estimating Regular Vine Copulae and Application to Financial Returns." *Computational Statistics & Data Analysis*, **59**, 52–69. doi:10.1016/j.csda.2012.08.010.

Embrechts P, Lindskog F, Mcneil A (2003). "Chapter 8 – Modelling Dependence with Copulas and Applications to Risk Management." In ST Rachev (ed.), *Handbook of Heavy Tailed Distributions in Finance*, volume 1 of *Handbooks in Finance*, pp. 329–384. North-Holland, Amsterdam. doi:10.1016/B978-044450896-6.50010-8.

Flora DB, Curran PJ (2004). "An Empirical Evaluation of Alternative Methods of Estimation for Confirmatory Factor Analysis with Ordinal Data." *Psychological Methods*, **9**(4), 466–491. doi:10.1037/1082-989x.9.4.466.

Foldnes N, Grønneberg S (2015). "How General is the Vale–Maurelli Simulation Approach?" *Psychometrika*, **80**(4), 1066–1083. doi:10.1007/s11336-014-9414-0.

Foldnes N, Grønneberg S (2017). "The Asymptotic Covariance Matrix and Its Use in Simulation Studies." *Structural Equation Modeling: A Multidisciplinary Journal*, **24**(6), 881–896. doi:10.1080/10705511.2017.1341320.

Foldnes N, Grønneberg S (2018). "Approximating Test Statistics Using Eigenvalue Block Averaging." *Structural Equation Modeling: A Multidisciplinary Journal*, **25**(1), 101–114. doi:10.1080/10705511.2017.1373021.

Foldnes N, Grønneberg S (2019a). "On Identification and Non-Normal Simulation in Ordinal Covariance and Item Response Models." *Psychometrika*, **84**(4), 1000–1017. doi:10.1007/s11336-019-09688-z.

Foldnes N, Grønneberg S (2019b). "Pernicious Polychorics: The Impact and Detection of Underlying Non-Normality." *Structural Equation Modeling: A Multidisciplinary Journal*, **27**(4), 525–543. doi:10.1080/10705511.2019.1673168.

Foldnes N, Grønneberg S (2021). ***discnorm**: Test for Discretized Normality in Ordinal Data.* R package version 0.1.1, URL `https://CRAN.R-project.org/package=discnorm`.

Foldnes N, Grønneberg S (2022a). ***covsim**: VITA, IG and PLSIM Simulation for Given Covariance and Marginals.* R package version 1.0.0, URL `https://CRAN.R-project.org/package=covsim`.

Foldnes N, Grønneberg S (2022b). "The Sensitivity of Structural Equation Modeling with Ordinal Data to Underlying Non-Normality and Observed Distributional Forms." *Psychological Methods*. `doi:10.1037/met0000385`. Forthcoming.

Foldnes N, Olsson UH (2015). "Correcting Too Much or Too Little? The Performance of Three Chi-Square Corrections." *Multivariate Behavioral Research*, **50**(5), 533–543. `doi:10.1080/00273171.2015.1036964`.

Foldnes N, Olsson UH (2016). "A Simple Simulation Technique for Nonnormal Data with Prespecified Skewness, Kurtosis, and Covariance Matrix." *Multivariate Behavioral Research*, **51**(2–3), 207–219. `doi:10.1080/00273171.2015.1133274`.

Fouladi RT (2000). "Performance of Modified Test Statistics in Covariance and Correlation Structure Analysis under Conditions of Multivariate Nonnormality." *Structural Equation Modeling: A Multidisciplinary Journal*, **7**(3), 356–410. `doi:10.1207/s15328007sem0703_2`.

Frees EW, Valdez EA (1998). "Understanding Relationships Using Copulas." *North American Actuarial Journal*, **2**(1), 1–25. `doi:10.1080/10920277.1998.10595667`.

Genest C, Favre AC (2007). "Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask." *Journal of Hydrologic Engineering*, **12**(4), 347–368. `doi:10.1061/(asce)1084-0699(2007)12:4(347)`.

Grimm KJ, Widaman KF (2010). "Residual Structures in Latent Growth Curve Modeling." *Structural Equation Modeling: A Multidisciplinary Journal*, **17**(3), 424–442. `doi:10.1080/10705511.2010.489006`.

Grønneberg S, Foldnes N (2017). "Covariance Model Simulation Using Regular Vines." *Psychometrika*, **82**(4), 1035–1051. `doi:10.1007/s11336-017-9569-6`.

Grønneberg S, Foldnes N (2019a). "A Problem with Discretizing Vale-Maurelli in Simulation Studies." *Psychometrika*, **84**(2), 554–561. `doi:10.1007/s11336-019-09663-8`.

Grønneberg S, Foldnes N (2019b). "Testing Model Fit by Bootstrap Selection." *Structural Equation Modeling: A Multidisciplinary Journal*, **26**(2), 182–190. `doi:10.1080/10705511.2018.1503543`.

Grønneberg S, Hjort NL (2014). "The Copula Information Criteria." *Scandinavian Journal of Statistics*, **41**(2), 436–459. `doi:10.1111/sjos.12042`.

Hofert M, Kojadinovic I, Maechler M, Yan J (2020). ***copula**: Multivariate Dependence with Copulas.* R package version 1.0-1, URL `http://CRAN.R-project.org/package=copula`.

Höffding W (1940). "Masstabinvariante Korrelationstheorie." In *Schriften des Mathematischen Instituts und Instituts für Angewandte Mathematik der Universität Berlin*, volume 5, pp. 181–233.

Joe H (1996). "Families of $m$-Variate Distributions with Given Margins and $m(m-1)/2$ Bivariate Dependence Parameters." In *Distributions with Fixed Marginals and Related Topics*, IMS Lecture Notes – Monograph Series, pp. 120–141.

Joe H (1997). *Multivariate Models and Multivariate Dependence Concepts*, volume 73. Chapman & Hall/CRC. doi:10.1201/9780367803896.

Joe H (2014). *Dependence Modeling with Copulas*. Chapman & Hall/CRC. doi:10.1201/b17116.

Jöreskog KG (1967). "Some Contributions to Maximum Likelihood Factor Analysis." *Psychometrika*, **32**(4), 443–482. doi:10.1007/bf02289658.

Jöreskog KG, Sörbom D (2006). "**LISREL** Version 8.8."

Kallenberg O (2002). *Foundations of Modern Probability*. 2nd edition. Springer-Verlag. doi:10.1007/978-1-4757-4015-8.

Kurowicka D, Joe H (eds.) (2011). *Dependence Modeling: Vine Copula Handbook*. World Scientific. doi:10.1142/7699.

Laenen A, Alonso A, Molenberghs G, Vangeneugden T (2009). "Reliability of a Longitudinal Sequence of Scale Ratings." *Psychometrika*, **74**(1), 49. doi:10.1007/s11336-008-9079-7.

Li CH (2016). "The Performance of ML, DWLS, and ULS Estimation with Robust Corrections in Structural Equation Models with Ordinal Variables." *Psychological Methods*, **21**(3), 369–387. doi:10.1037/met0000093.

Mair P, Satorra A, Bentler PM (2012). "Generating Nonnormal Multivariate Data Using Copulas: Applications to SEM." *Multivariate Behavioral Research*, **47**(4), 547–565. doi:10.1080/00273171.2012.692629.

Marcoulides KM (2019). "Reliability Estimation in Longitudinal Studies Using Latent Growth Curve Modeling." *Measurement: Interdisciplinary Research and Perspectives*, **17**(2), 67–77. doi:10.1080/15366367.2018.1522169.

Marcoulides KM, Foldnes N, Grønneberg S (2020). "Assessing Model Fit in Structural Equation Modeling Using Appropriate Test Statistics." *Structural Equation Modeling: A Multidisciplinary Journal*, **27**(3), 369–379. doi:10.1080/10705511.2019.1647785.

Micceri T (1989). "The Unicorn, the Normal Curve, and Other Improbable Creatures." *Psychological Bulletin*, **105**(1), 156–166. doi:10.1037/0033-2909.105.1.156.

Muthén B (1984). "A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators." *Psychometrika*, **49**(1), 115–132. doi:10.1007/bf02294210.

Nagler T, Vatter T (2021). ***rvinecopulib**: High Performance Algorithms for Vine Copula Modeling.* R package version 0.6.1.1.1, URL `https://CRAN.R-project.org/package=rvinecopulib`.

Nelsen RB (2007). *An Introduction to Copulas.* Springer-Verlag. `doi:10.1007/0-387-28678-0`.

Pearson K (1895). "Contributions to the Mathematical Theory of Evolution.—II. Skew Variation in Homogeneous Material." *Philosophical Transactions of the Royal Society of London. A*, **186**, 343–414. `doi:10.1098/rsta.1895.0010`.

Pornprasertmanit S, Miller P, Schoemann A, Jorgensen TD (2021). ***simsem**: SIMulated Structural Equation Modeling.* R package version 0.5-16, URL `https://CRAN.R-project.org/package=simsem`.

Qu W, Liu H, Zhang Z (2019). "A Method of Generating Multivariate Non-Normal Random Numbers with Desired Multivariate Skewness and Kurtosis." *Behavior Research Methods*, **52**(3), 939–946. `doi:10.3758/s13428-019-01291-5`.

Qu W, Zhang Z (2020). ***mnonr**: A Generator of Multivariate Non-Normal Random Numbers.* R package version 1.0.0, URL `https://CRAN.R-project.org/package=mnonr`.

Quiroga AM (1994). *Studies of the Polychoric Correlation and Other Correlation Measures for Ordinal Variables.* Ph.D. thesis, Uppsala University.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org/`.

Rhemtulla M, Brosseau-Liard PÉ, Savalei V (2012). "When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods under Suboptimal Conditions." *Psychological Methods*, **17**(3), 354–373. `doi:10.1037/a0029315`.

Rice JA (2006). *Mathematical Statistics and Data Analysis.* Cengage Learning.

Rosseel Y (2012). "**lavaan**: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. `doi:10.18637/jss.v048.i02`.

Rüschendorf L (2009). "On the Distributional Transform, Sklar's Theorem, and the Empirical Copula Process." *Journal of Statistical Planning and Inference*, **139**(11), 3921–3927. `doi:10.1016/j.jspi.2009.05.030`.

Ruscio J, Kaczetow W (2008). "Simulating Multivariate Nonnormal Data Using an Iterative Algorithm." *Multivariate Behavioral Research*, **43**(3), 355–381. `doi:10.1080/00273170802285693`.

Satorra A, Bentler PM (1988). "Scaling Corrections for Statistics in Covariance Structure Analysis." *UCLA Statistics Series 2*, University of California at Los Angeles, Department of Psychology, Los Angeles.

Schepsmeier U, Stoeber J, Brechmann EC, Graeler B, Nagler T, Erhardt T (2021). ***VineCopula**: Statistical Inference of Vine Copulas.* R package version 2.4.3, URL `https://CRAN.R-project.org/package=VineCopula`.

Shapiro A (1983). "Asymptotic Distribution Theory in the Analysis of Covariance Structures." *South African Statistical Journal*, **17**(1), 33–81.

Shorack GR, Wellner JA (2009). *Empirical Processes with Applications to Statistics*, volume 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia. doi:10.1137/1.9780898719017. Originally published in 1986 by John Wiley & Sons Inc., New York.

Sklar M (1959). *Fonctions de Répartition à n Dimensions et Leurs Marges*. Publications de l'Institut Statistique de l'Université de Paris 8.

Takane Y, de Leeuw J (1987). "On the Relationship Between Item Response Theory and Factor Analysis of Discretized Variables." *Psychometrika*, **52**(3), 393–408. doi:10.1007/bf02294363.

Tarka P (2018). "An Overview of Structural Equation Modeling: Its Beginnings, Historical Development, Usefulness and Controversies in the Social Sciences." *Quality & Quantity*, **52**(1), 313–354. doi:10.1007/s11135-017-0469-8.

Touloumis A (2016). "Simulating Correlated Binary and Multinomial Responses Under Marginal Model Specification: The **SimCorMultRes** Package." *The R Journal*, **8**(2), 79–91. doi:10.32614/rj-2016-034.

Vale CD, Maurelli VA (1983). "Simulating Multivariate Nonnormal Distributions." *Psychometrika*, **48**(3), 465–471. doi:10.1007/bf02293687.

Van De Schoot R, Sijbrandij M, Winter SD, Depaoli S, Vermunt JK (2017). "The GRoLTS-Checklist: Guidelines for Reporting on Latent Trajectory Studies." *Structural Equation Modeling: A Multidisciplinary Journal*, **24**(3), 451–467. doi:10.1080/10705511.2016.1247646.

Wu H, Lin J (2016). "A Scaled $F$ Distribution as an Approximation to the Distribution of Test Statistics in Covariance Structure Analysis." *Structural Equation Modeling: A Multidisciplinary Journal*, **23**(3), 409–421. doi:10.1080/10705511.2015.1057733.

Yan J (2007). "Enjoy the Joy of Copulas: With a Package **copula**." *Journal of Statistical Software*, **21**(4), 1–21. doi:10.18637/jss.v021.i04.

# A. How simulation is done on a computer

## A.1. How uni- and bivariate simulations are performed on a computer

Throughout the paper, we assume that the marginals of the distribution we simulate from are continuous. In SEM, this is mostly without loss of generality, as ordinal variables are usually modeled as discretizations of a continuous random vector, see Section 6.2. Note that this also applies to a large class of IRT models (see, e.g., Foldnes and Grønneberg 2019a; Takane and de Leeuw 1987).

### *Univariate simulation*

We first review a standard method to simulate from a continuous univariate distribution with cumulative distribution function $F_1$. This is covered in most standard statistics text books (see, e.g., Rice 2006, Proposition D, Section 2.3).

We assume further that $F_1$ is strictly increasing, and therefore has an inverse $F_1^{-1}$. Since $F_1$ is a continuous distribution function, it will be continuous and increasing, but not necessarily strictly increasing (i.e., there may be flat regions), necessitating the use of more complex arguments, such as those in Chapter 1 of Shorack and Wellner (2009). We will throughout the paper pedagogically assume such strict monotonicity, and thereby avoiding such complex arguments.

Recall that the inverse function $F_1^{-1}$ is defined as the solution to the equation $F_1(x) = u$ with respect to $x$, and where $0 < u < 1$. Clearly, this solution depends on $u$, and we therefore denote the solution as a function of $u$, that is, $F_1^{-1}(u)$. Since $F_1^{-1}(u) = x$ where $F_1(x) = u$ we get that $F(F^{-1}(u)) = F(x) = u$. We may compute $F_1^{-1}(u)$ by solving

$$F_1(x) - u = 0 \tag{2}$$

for $u$. Since $F_1$ is increasing and continuous, with $F_1(-\infty) = 0$ and $F_1(\infty) = 1$, (2) has a single solution, which can be found either analytically, or using any standard root finding procedure.

Now let $U$ be a univariate random variable with the uniform distribution on $[0, 1]$, denoted by $U \sim U[0, 1]$. This means that $P(U \leq x) = xI\{0 \leq x \leq 1\} + I\{x > 1\}$ where $I\{A\}$ is the indicator function of $A$ which is 1 if $A$ is true, and zero otherwise. We may then let

$$X = F_1^{-1}(U).$$

The distribution of $X$ is $F_1$. To see this, we start with $P(X \leq x) = P(F_1^{-1}(U) \leq x)$. Applying $F_1$ on both sides of the inequality is allowed since $F_1$ is increasing. Since $F(F_1^{-1}(U)) = U$, this gives

$$\begin{aligned} P(F_1^{-1}(U) \leq x) &= P(F_1(F_1^{-1}(U)) \leq F_1(x)) = P(U \leq F_1(x)) \\ &= F_1(x)I\{0 \leq F_1(x) \leq 1\} + I\{F_1(x) > 1\}. \end{aligned}$$

Since $F_1$ is a cumulative distribution function, we have $0 \leq F_1(x) \leq 1$, and therefore the first indicator function is always one, while the second is always zero. Therefore, we conclude that $P(X \leq x) = F_1(x)$ as required.

*Imposing required marginals on copula-distributions*

Let us now consider the more complex bivariate case, which is not as well-known as the univariate case. Firstly, let us assume that we are able to simulate from a copula $C$. That is, suppose $(U_1, U_2)^\top \sim C$, meaning

$$P(U_1 \leq u_1, U_2 \leq u_2) = C(u_1, u_2). \tag{3}$$

Recall that a copula has uniform marginals. This means that $U_1 \sim U[0,1]$ and $U_2 \sim U[0,1]$. Therefore, following the argument above, we may let $X_1 = F_1^{-1}(U_1)$ and $X_2 = F_2^{-1}(U_2)$, and see that the marginal distribution of $X_1$ is $F_1$ and the marginal distribution of $X_2$ is $F_2$. This means $P(X_i \leq x_i) = F_i(x_i)$ for $i = 1, 2$, and does not say anything about the dependence between $X_1$ and $X_2$. We now show that $(X_1, X_2)$ has the distribution as in (1), i.e., $P(X_1 \leq x_1, X_2 \leq X_2) = C(F_1(x_1), F_2(x_2))$. Using the definition of $X_1, X_2$ and applying the functions $F_1$ and $F_2$ respectively on the two inequalities, we have

$$\begin{aligned} P(X_1 \leq x_1, X_2 \leq x_2) &= P(F_1^{-1}(U_1) \leq x_1, F_2^{-1}(U_2) \leq x_2) \\ &= P(U_1 \leq F_1(x_1), U_2 \leq F_2(x_2)) \\ &= C(F_1(x_1), F_2(x_2)) \end{aligned}$$

where the last equality uses (3).

*Bivariate copula simulation*

Let us now review how to simulate $(U_1, U_2)$ from $C$. We follow a general method described in Nelsen (2007, Section 2.9), which uses the so-called multivariate quantile transform. In the univariate case, the central step in simulating from $F_1$ was to compute the function $F_1^{-1}$. In the bivariate case, the central step is to compute a function $h_{u_1}^{-1}(u_2)$, which will later be shown to be the inverse of the distribution function of $U_2$ conditional on $U_1$. After having written code which can evaluate this function, the details of which will be covered shortly, we may simulate from $C$ as follows: Let $U_1 \sim U[0,1]$ and $V \sim U[0,1]$ be independent. Then set $U_2 = h_{U_1}^{-1}(V)$. The pair $(U_1, U_2)$ is then distributed according to the copula $C$ (Nelsen 2007, Section 2.9).

We now define $h_{u_1}^{-1}(u_2)$. For each value $0 < u_1 < 1$, let

$$h_{u_1}(u_2) = \frac{\partial}{\partial u_1} C(u_1, u_2).$$

Then, for each $0 < u_1 < 1$, the function $h_{u_1}^{-1}$ is the (generalized) inverse of $h_{u_1}(u_2)$ as a function of $u_2$. Recall that for a function $H(x)$, its generalized inverse is defined as $H^{-1}(y) = \inf\{x : H(x) > y\}$, a definition which agrees with the standard inverse when $H$ is invertible. As in the univariate case, where it was simpler to work with the case when $F_1$ was continuous and strictly increasing, we will again assume that $h_{u_1}$ is continuous and strictly increasing for all $u_1$. We now show that this follows if $C$ has a density $c$ which is non-zero on $(0,1)^2$ and continuous in each coordinate. Both assumptions on $c$ are fulfilled for all popular copula classes, and will therefore be assumed also in the following. To see that these two assumptions imply that $h_{u_1}$ is continuous and strictly increasing for any $u_1$, notice that since

$$C(u_1, u_2) = \int_0^{u_1} \int_0^{u_2} c(x_1, x_2)\, dx_1 dx_2,$$

we have $h_{u_1}(u_2) = \int_0^{u_2} c(u_1, x_2)\, dx_2$ by the fundamental theorem of calculus. Since $c(u_1, u_2) > 0$, and since the function $x_2 \mapsto c(u_1, x_2)$ is continuous for a given $u_1$, the integral is strictly increasing and continuous in $u_2$ and hence $h_{u_1}(u_2)$ is strictly increasing and continuous in $u_2$. Further, we have $h_{u_1}(0) = 0$ and $h_{u_1}(1) = 1$, and $h_{u_1} : [0, 1] \mapsto [0, 1]$ is therefore a bijection for each $u_1$. That $h_{u_1}(0) = 0$ and $h_{u_1}(1) = 1$ can be seen from noticing that $h_{u_1}(0) = \frac{\partial}{\partial u_1} C(u_1, 0) = \frac{\partial}{\partial u_1} 0 = 0$ since $C(u_1, 0) = P(U_1 \leq u_1, U_2 \leq 0) = 0$, and that since $C(u_1, 1) = P(U_1 \leq u_1, U_2 \leq 1) = P(U_1 \leq u_1) = u_1$ we also have $h_{u_1}(1) = \frac{\partial}{\partial u_1} C(u_1, 1) = \frac{\partial}{\partial u_1} u_1 = 1$.

Finally, we show that $(U_1, U_2)^\top \sim C$. We have not found an elementary presentation of this result in the literature (see Rüschendorf 2009, Section 3, for an authoritative account), and include it for completeness and since the following argument is short, and will be generalized progressively in the following. For compactness, we assume $h_{u_1}(u_2)$ is invertible with respect to $u_2$ for all $0 < u_1 < 1$. We have

$$\begin{aligned}
P(U_1 \leq u_1, U_2 \leq u_2) &= P(U_1 \leq u_1, h_{U_1}^{-1}(V) \leq u_2) \\
&\overset{(a)}{=} P(U_1 \leq u_1, V \leq h_{U_1}(u_2)) \overset{(b)}{=} \mathsf{E}\mathsf{E}[I\{U_1 \leq u_1\} I\{V \leq h_{U_1}(u_2)\} | U_1] \\
&\overset{(c)}{=} \mathsf{E} I\{U_1 \leq u_1\} \mathsf{E}[I\{V \leq h_{U_1}(u_2)\} | U_1] \overset{(d)}{=} \mathsf{E} I\{U_1 \leq u_1\} h_{U_1}(u_2) \\
&= \int_0^1 I\{x_1 \leq u_1\} h_{x_1}(u_2)\, dx_1 = \int_0^{u_1} h_{x_1}(u_2)\, dx_1 = \int_0^{u_1} \frac{\partial}{\partial x_1} C(x_1, u_2)\, dx_1 \\
&\overset{(e)}{=} C(u_1, u_2) - C(0, u_2) \overset{(f)}{=} C(u_1, u_2)
\end{aligned}$$

Explanations: *(a)* Apply $h_{U_1}$ to both sides of the second inequality. *(b)* $P(A) = \mathsf{E} I\{A\}$. Double expectation. Also, $I\{A, B\} = I\{A\} I\{B\}$. *(c)*: Use $\mathsf{E}[H(X) Y | X] = H(X) \mathsf{E}[Y | X]$. *(d)*: Since $V$ is independent to $U_1$, we have $\mathsf{E}[I\{V \leq h_{U_1}(u_2)\} | U_1] = \mathsf{E}_V I\{V \leq h_{U_1}(u_2)\}$, where the expectation is with respect to $V$ only, and $U_1$ is fixed. Recalling that for a random variable $X$ with CDF $F_X$ we have $F_X(x) = P(X \leq x) = \mathsf{E} I\{X \leq x\}$ we have that $\mathsf{E}_V I\{V \leq h_{U_1}(u_2)\} = h_{U_1}(u_2)$ since $V$ is uniform on $[0, 1]$, i.e., $V$ has CDF $F_U(u) = u$ for $0 \leq u \leq 1$. *(e)* Fundamental theorem of calculus. *(f)* Since $0 \leq U_1 \leq 1$ and continuous we have $C(0, u_2) = P(U_1 \leq 0, U_2 \leq u_2) = P(U_1 = 0, U_2 \leq u_2) = 0$.

Practically speaking, we may therefore simulate from any copula by computing $h_{u_1}^{-1}(u_2)$, which requires the computation of a partial derivative and the inversion of a function. This may be done analytically in some cases, but in general, numerical approximation is required.

### *Identifying correlations from a bivariate distribution*

We now consider how we may identify a $\theta_0$ such that the distribution $F(x_1, x_2; \theta_0) = C(F_1(x_1), F_2(x_2); \theta_0)$ has a Pearson correlation $\rho$. The Höffding (1940) formula for correlation $\rho(\theta)$ gives

$$\begin{aligned}
\rho(\theta) = \mathrm{sd}(F_1)^{-1} \mathrm{sd}(F_2)^{-1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} & C(F_1(z_1), F_2(z_2); \theta) \\
& - F_1(z_1) F_2(z_2)\, dz_1 dz_2,
\end{aligned} \tag{4}$$

where $\mathrm{sd}(F_1), \mathrm{sd}(F_2)$ are the standard deviations of $F_1, F_2$. Most bivariate copula classes are such that $C(u_1, u_2; \theta)$ is increasing in $\theta$, a property fulfilled by all copulas cataloged in Section 5.1 of Joe (1997). This implies (Grønneberg and Foldnes 2017, Theorem 1) that $\rho(\theta)$ is an increasing function. It is therefore easy to solve for $\theta_0$ via numerical root finding methods. The function $\rho(\theta)$ then has to be evaluated through numerical integration.

## A.2. How trivariate vine simulations are performed on a computer

We here provide more technical details on how to simulate from the three-dimensional vine distribution used as an example in the main text. As is generally the case, we only need to simulate from the vine copula (which by definition has uniform marginals), as the marginals are easily transformed to any desired marginal distributions in the same way as explained in the univariate and bivariate case, i.e., by applying the quantile functions of the marginals to each of the coordinates of the simulated vector from a copula. When this is done, the resulting distribution has the desired marginals, and the same copula as before transformation. That is, if $C$ is a $d$-dimensional copula, meaning $C$ is a $d$-dimensional cumulative distribution function with uniform marginals, and $(U_1, U_2, \dots, U_d)^\top \sim C$, then $X = (F_1^{-1}(U_1), F_2^{-1}(U_2), \dots, F_d^{-1}(U_d))$ will have a distribution of the form

$$F(x_1, x_2, \dots, x_d) = C(F_1(x_1), F_2(x_d), \dots, F_d(x_d)),$$

meaning $X$ has marginals $F_1, F_2, \dots, F_d$ and $C$ as its copula.

Let us first consider how to simulate from a three-dimensional copula in general. Constructing multivariate distributions can be difficult, and vines provide a general construction which is often useful. A useful feature of this class is that some of the properties of the resulting distribution are well suited for computation, and in spite of the flexibility and simplicity of constructing vine distributions, simulating from them is straightforward.

We first consider how to simulate from a general three-dimensional copula $(U_1, U_2, U_3)^\top \sim C$, which does not have to be a vine distribution. We first simulate $(U_1, U_2)$ from the bivariate copula $C_{1,2}$ as described in the bivariate section above. Recall that $C_{1,2}$ can be easily computed from $C$, since $0 \leq U_3 \leq 1$ implies that $C_{1,2}(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2) = P(U_1 \leq u_1, U_2 \leq u_2, U_3 \leq 1) = C(u_1, u_2, 1)$. We may now simulate $U_3$ from the distribution of $(U_1, U_2, U_3)$ after "conditioning away" the values of $U_1, U_2$. Again we may use the multivariate quantile transform. We simulate $V$ which is uniform on $[0, 1]$ and independent from previously generated variables, and then let

$$U_3 = h^{-1}_{U_1,U_2}(V),$$

where $h^{-1}_{U_1,U_2}$ is the generalized inverse of $h_{U_1,U_2}$, which is the distribution of $U_3$ conditional on $(U_1, U_2)$. That is,

$$h_{u_1,u_2}(u_3) = C_{3|12}(u_3|u_1, u_2).$$

Conditional distributions are conceptually complex, and will only be covered superficially here. While we will give some more needed technical details for conditional distributions in the next subsection, we follow common text-book treatments (e.g., Rice 2006), and use the fact that if $C$ has a density $c_{1,2,3}$, which we will assume, $C_{3|12}$ has density given by

$$c_{3|12}(u_3|u_1, u_2) = \frac{c_{1,2,3}(u_1, u_2, u_3)}{c_{1,2}(u_1, u_2)},$$

where

$$c_{1,2}(u_1, u_2) = \int_0^1 c_{1,2,3}(u_1, u_2, x_3)\, dx_3.$$

We then have that

$$\begin{aligned}
C_{3|12}(u_3|u_1,u_2) &= \int_0^{u_3} c_{3|12}(u_3|u_1,u_2)\,dx_3 = \int_0^{u_3} \frac{c_{1,2,3}(u_1,u_2,x_3)}{c_{1,2}(u_1,u_2)}\,dx_3 \\
&= \frac{\int_0^{u_3} c(u_1,u_2,x_3)\,dx_3}{c_{1,2}(u_1,u_2)}.
\end{aligned}$$

As in the bivariate case, $h_{u_1,u_2}(u_3)$ is seen to be invertible if we assume that $c$ is continuous and non-zero on $(0,1)^3$, and generalized inverses are not needed when computing $U_3$. If a simple formula for the copula CDF is available, which we note is often not the case, we may avoid integration when computing $h_{u_1,u_2}$, since

$$C_{3|12}(u_3|u_1,u_2) = \frac{\int_0^{u_3} c(u_1,u_2,x_3)\,dx_3}{c_{1,2}(u_1,u_2)} = \frac{\frac{\partial^2}{\partial u_1 \partial u_2} C_{1,2,3}(u_1,u_2,u_3)}{\frac{\partial^2}{\partial u_1 \partial u_2} C_{1,2,3}(u_1,u_2,1)}. \tag{5}$$

When only the joint density of $C$ is available, integration is in general needed for computing $C_{3|12}$, and this is achieved by numerical approximations.

Following this recipe gives $(U_1,U_2,U_3)^\top \sim C_{1,2,3}$ by a similar argument as in the bivariate case: Again we assume $h_{u_1,u_2}(u_3)$ is invertible as a function of $u_3$ for all $0 < u_1, u_2 < 1$. We then have

$$\begin{aligned}
P(U_1 \le u_1, U_2 \le u_2, U_3 \le u_3) &= \mathsf{E} I\{U_1 \le u_1, U_2 \le u_2\} I\{h^{-1}_{U_1,U_2}(V) \le u_3\} \\
&= \mathsf{E}\mathsf{E}[I\{U_1 \le u_1, U_2 \le u_2\} I\{V \le h_{U_1,U_2}(u_3)\}|U_1,U_2] \\
&= \mathsf{E} I\{U_1 \le u_1, U_2 \le u_2\} \mathsf{E}[I\{V \le h_{U_1,U_2}(u_3)\}|U_1,U_2] \\
&\overset{(a)}{=} \mathsf{E} I\{U_1 \le u_1, U_2 \le u_2\} \mathsf{E}_V[I\{V \le h_{U_1,U_2}(u_3)\}] \\
&\overset{(b)}{=} \mathsf{E} I\{U_1 \le u_1, U_2 \le u_2\} h_{U_1,U_2}(u_3) \\
&= \int_0^{u_1}\int_0^{u_2} c_{12}(x_1,x_2) h_{x_1,x_2}(u_3)\,dx_1 dx_2 \\
&= \int_0^{u_1}\int_0^{u_2} \frac{\partial^2}{\partial x_1 \partial x_2} C_{1,2,3}(x_1,x_2,u_3)\,dx_1 dx_2 \\
&= C_{1,2,3}(u_1,u_2,u_3).
\end{aligned}$$

Explanations: *(a)* $V$ is independent to $U_1, U_2$. *(b)* $V$ is uniform.

Let us now apply this to three-dimensional regular vines. The idea behind vines is described in Joe (1996) and Joe (2014, Sections 3.8 and 3.9), and is based on expressing cumulative distribution functions as mixtures of conditional distributions. The motivation for its construction will be sketched in the next subsection, and we here simply state the distribution for a trivariate copula in terms of its CDF.

The three-dimensional vine copula illustrated in Figure 4 has cumulative distribution

$$C_{1,2,3}(u_1,u_2,u_3) = \int_0^{u_2} C_{1,3;2}(C_{1|2}(u_1|z_2), C_{3|2}(u_3|z_2))\,dz_2, \tag{6}$$

where $C_{1,3;2}$ is a chosen bivariate copula to bind marginals 1 and 3 together, when given 2, and where

$$C_{1|2}(u_1, u_2) = \frac{\partial}{\partial u_2} C_{1,2}(u_1, u_2) = \int_0^{u_2} c_{1,2}(u_1, x_2)\, dx_2,$$
$$C_{3|2}(u_3, u_2) = \frac{\partial}{\partial u_2} C_{3,2}(u_3, u_2) = \int_0^{u_2} c_{3,2}(u_3, x_2)\, dx_2,$$

with $C_{1,2}$ and $C_{3,2}$ the chosen copulas, directly giving the copula of marginals $1, 2$ and $3, 2$ respectively. Note that $C_{1,3;2}$ is a standard bivariate copula, and does not depend on the value $x_2$ being integrated over in the above display. This is an important point which will be discussed further in the next subsection.

There is an important distinction between $C_{1,3|2}$, which is the conditional distribution of $(U_1, U_3)$ given $U_2$, and $C_{1,3;2}$, which as we will discuss later is the *copula* of $C_{1,3|2}$. The objects $C_{1,3|2}$ and $C_{1,3;2}$ need not be the same, and while $C_{1,2,3}$ has all uniform marginals, the marginals of the conditional distribution $C_{1,3|2}$ need *not* be uniform, which in turn implies that $C_{1,3|2}$ need not be a copula. We will return to this issue in the following. In order to further separate the two further, we will keep the $C$ notation for functions such as $C_{1,3;2}$ – since these *are* actually copulas, but rather write $F_{1,3|2}$ to refer to the conditional distribution of $(U_1, U_3)$ given $U_2$. That is, we will from now on write $F_{1,3|2} = C_{1,3|2}$.

Consider now how to simulate from this vine. Since $(U_1, U_2)$ is simulated from $C_{1,2}$ which is directly specified in the lowest tree, its simulation procedure follows from the already described bivariate case. We now need to compute $F_{3|12}$. Recalling (5), we have

$$\begin{aligned} F_{3|12}(u_3|u_1, u_2) &= \frac{\frac{\partial^2}{\partial u_1 \partial u_2} C_{1,2,3}(u_1, u_2, u_3)}{c_{12}(u_1, u_2)} \\ &= \frac{\frac{\partial^2}{\partial u_1 \partial u_2} \int_0^{u_2} C_{1,3;2}(C_{1|2}(u_1|z_2), C_{3|2}(u_3|z_2))\, dz_2}{c_{12}(u_1, u_2)} \\ &= \frac{\frac{\partial}{\partial u_1} C_{1,3;2}(C_{1|2}(u_1|u_2), C_{3|2}(u_3|u_2))}{c_{12}(u_1, u_2)}. \end{aligned}$$

A notable feature is that this expression only depends on bivariate distributions, which are usually computationally well-behaved.

## A.3. More details on the vine construction in the trivariate case

We here provide a sketch of the vine construction of Joe (1996). We are unaware of an elementary presentation of this material in the literature, and presentations such as those in Joe (1996); Bedford and Cooke (2002); Joe (2014) require considerable technical training to read. We therefore include an elementary presentation of this material here, restricted to the trivariate case.

Using operations similar to the derivation on the validity of the general trivariate simulation method, we see that for any trivariate continuous copula $C$, we have for variables

$(U_1, U_2, U_3)^\top \sim C$ that

$$\begin{aligned}
C(u_1, u_2, u_3) &= P(U_1 \leq u_1, U_2 \leq u_2, U_3 \leq u_3) \\
&= \mathsf{E} I\{U_1 \leq u_1, U_2 \leq u_2, U_3 \leq u_3\} \\
&= \mathsf{E} I\{U_2 \leq u_2\} I\{U_1 \leq u_1, U_3 \leq u_3\} \\
&= \mathsf{E}\mathsf{E}[I\{U_2 \leq u_2\} I\{U_1 \leq u_1, U_3 \leq u_3\} | U_2] \\
&= \mathsf{E} I\{U_2 \leq u_2\} \mathsf{E}[I\{U_1 \leq u_1, U_3 \leq u_3\} | U_2] \\
&= \mathsf{E} I\{U_2 \leq u_2\} P(U_1 \leq u_1, U_3 \leq u_3 | U_2) \\
&= \int_0^1 I\{x_2 \leq u_2\} P(U_1 \leq u_1, U_3 \leq u_3 | U_2 = x_2)\, dx_2 \\
&= \int_0^{u_2} P(U_1 \leq u_1, U_3 \leq u_3 | U_2 = x_2)\, dx_2.
\end{aligned}$$

This calculation provides an expansion of the full distribution of $(U_1, U_2, U_3)$ in terms of the conditional distribution of $(U_1, U_3)$ given $U_2$. This conditional bivariate distribution $F_{1,3|2}(u_1, u_3|x_2) = P(U_1 \leq u_1, U_3 \leq u_3|U_2 = x_2)$ has marginals $F_{1|2}(u_1|x_3)$ and $F_{3|2}(u_3|x_2)$, which can be derived using properties of conditional distributions. A non-rigorous heuristic argument for the formula for $F_{1|2}(u_1|x_2)$ is that

$$\begin{aligned}
F_{1|2}(u_1|x_2) &= P(U_1 \leq u_1 | U_2 = x_2) \\
&= \lim_{h \to 0} \frac{P(U_1 \leq u_1, x_2 \leq U_2 \leq x_2 + h)}{P(x_2 \leq U_2 \leq x_2 + h)} \\
&= \lim_{h \to 0} \frac{C_{1,2}(u_1, u_2 + h) - C_{1,2}(u_1, u_2)}{x_2 + h - x_2} \\
&= \lim_{h \to 0} \frac{C_{1,2}(u_1, u_2 + h) - C_{1,2}(u_1, u_2)}{h} \\
&= \frac{\partial}{\partial u_2} C_{1,2}(u_1, u_2),
\end{aligned}$$

using the uniformity of $U_2$. Similarly, $F_{3|2}(u_3|x_2) = \frac{\partial}{\partial u_2} C_{3,2}(u_3, u_2)$. A formal argument justifying the formulas for $F_{1|2}$ and $F_{3|2}$ requires the general and rather complex mathematical framework of conditional probability, as developed by Kolmogorov, see Kallenberg (2002). A nice feature following from our focus on simulation is that an alternative justification for the formula for conditional distributions is provided by its successful application in simulation.

Sklar's theorem applied for each given $x_2$ value to the conditional distribution

$$F_{1,3|2}(u_1, u_3|x_2) = P(U_1 \leq u_1, U_3 \leq u_3|U_2 = x_2)$$

shows that there is a class of copulas $C_{13;2}(u_1, u_3; x_2)$ varying with $x_2$, which is such that

$$F_{1,3|2}(u_1, u_3|x_2) = C_{13;2}(F_{1|2}(u_1|x_2), F_{3|2}(u_3|x_2); x_2).$$

Using the formulas we identified for $F_{1|2}(u_1|x_2), F_{3|2}(u_3|x_2)$ and recalling that we started with an expansion for $C(u_1, u_2, u_3)$, we have shown that

$$C(u_1, u_2, u_3) = \int_0^{u_2} C_{13;2}\left(\frac{\partial}{\partial u_2} C_{1,2}(u_1, u_2), \frac{\partial}{\partial u_2} C_{3,2}(u_3, u_2); x_2\right) dx_2, \tag{7}$$

which holds in general. This expression can also be used to construct multivariate distributions from bivariate distributions: Based on bivariate copulas $C_{1,2}$ and $C_{3,2}$ we may compute $\frac{\partial}{\partial u_2} C_{1,2}(u_1, u_2)$, $\frac{\partial}{\partial u_2} C_{3,2}(u_3, u_2)$, and they may be combined using a family of copulas $C_{13;2}(u_1, u_3; x_2)$ for every $x_2$. For each $x_2$, the Sklar theorem implies that

$$C_{13;2}\left(\frac{\partial}{\partial u_2} C_{1,2}(u_1, u_2), \frac{\partial}{\partial u_2} C_{3,2}(u_3, u_2); x_2\right)$$

is a proper distribution. However, the family of copulas $C_{13;2}(u_1, u_3; x_2)$ has to be linked together via their $x_2$ dependence in such that the resulting $C$ in (7) is a proper CDF. This may be challenging, and does not have a simple solution.

The vine copula construction assumes that the family $C_{13;2}(u_1, u_3; x_2)$ is constant in $x_2$, i.e., does not depend on $x_2$. This is known as the simplifying assumption (Joe 2014). We therefore write

$$C_{13;2}(u_1, u_3; x_2) = C_{13;2}(u_1, u_3), \tag{8}$$

and see that we re-gain the vine CDF of (6). Since $C_{13;2}(u_1, u_3)$ does not vary with $x_2$, the combination from (6) always results in a valid CDF, as may be seen as follows. We may consider the algorithm for simulating from $C_{1,2,3}$. After having simulated from $(U_1, U_2)$ using previously described bivariate techniques, we define $U_3 = h^{-1}_{U_1,U_2}(V)$ from an independent $V \sim U[0, 1]$. Clearly, $U_3$ is a random variable, and by the above argument, the joint distribution of $(U_1, U_2, U_3)$ is precisely $C_{1,2,3}$ from (6), and hence $C_{1,2,3}$ is indeed a valid distribution function since it is the CDF of a random vector. By (7) the constructed distribution has $C_{13;2}$ as the copula of the conditional distribution of $(U_1, U_3)$ given $U_2$.

## A.4. The density of a four-dimensional vine

In the main text, we gave the density of the three-dimensional vine in Figure 4 without a complete technical description. We here rectify this by deriving the density of the more general four-dimensional vine as depicted in Figure 10, and sketch how to form such densities in general. How to simulate from this four-dimensional vine will be the topic of the next section. Our discussion of this four-dimensional example ought to be sufficient to prepare the reader to understand general descriptions of simulation in, e.g., Dissmann, Brechmann, Czado, and Kurowicka (2013); Joe (2014), as well as fully understanding the vine based VITA simulation methodology of Grønneberg and Foldnes (2017).

The copula density of a vine is found by multiplying all copulas that are chosen as edge copulas. These copulas are evaluated at rather specific points, which will be discussed in the following. The edge copulas are the copulas of bivariate conditional distributions from the resulting full copula, with conditioning set indicated by the edge names. For example, the top-most edge connects $(U_1, U_4)$ and conditions on $(U_2, U_3)$, and represents the copula of $F_{1,4|2,3}$. Its contribution to the full density therefore includes a multiplicative factor $c_{1,4;2,4}$, where the use of a semi-colon indicates that this is the copula of a conditional distribution. As explained in the previous section, we write all conditional distributions of $c$ such as the actual conditional distribution (here, a density) of $(U_1, U_4)$ conditional on $(U_2, U_3)$ using the notation $f_{1,4|2,3}$ for the density and $F_{1,3|2,3}$ for the CDF.

Conditional marginals are the key to the general description of writing down the density of a vine copula based on its vine, such as Figure 10, as they are included in the multiplicative contribution from each edge. For any edge on the vine, the edge may be denoted by $(i, j|\mathbf{v})$

where $i, j$ are the marginals connected by this edge, and $\mathbf{v}$ may contain several indices (or none, as is the case at the lowest tree) which are conditioned on. For example, the top-most tree in Figure 10 contains only one edge, where $i = 1, j = 4$ and $\mathbf{v} = \{2, 3\}$. In contrast, the second tree in Figure 10 contains two edges. For the left-most edge, we have $i = 1, j = 3$ and $\mathbf{v} = \{2\}$. For the right-most edge, we have $i = 2, j = 4$ and $\mathbf{v} = \{3\}$. For the lowest tree, we do not condition on any indices, and so $\mathbf{v}$ is always empty. Going from left to right, the first edge has $i = 1, j = 2$, the next edge has $i = 2, j = 3$, and the final edge has $i = 3, j = 4$.

The multiplicative contribution of every edge $i, j|\mathbf{v}$ is $c_{i,j|\mathbf{v}}(F_{i|\mathbf{v}}(u_i|u_\mathbf{v}), F_{j|\mathbf{v}}(u_j|u_\mathbf{v}))$, where $u_\mathbf{v} = (u_k : k \in \mathbf{v})$, and where $F_{i|\mathbf{v}}(u_i|u_\mathbf{v})$ is the conditional cumulative distribution of $U_i$ given $\{U_k : k \in \mathbf{v}\}$.

When $\mathbf{v}$ is empty, $F_{i|\mathbf{v}}(u_i|u_\mathbf{v})$ is the actual cumulative distribution of $U_i$, which is uniform since $c$ is a copula. In these cases, we have $F_{i|\mathbf{v}}(u_i|u_\mathbf{v}) = u_i$. Therefore, the contributions of the lowest tree are simply $c_{1,2}(u_1, u_2)c_{2,3}(u_2, u_3)c_{3,4}(u_3, u_4)$.

Combining this description for the vine in Figure 10, we see that

$$\begin{aligned} c_{1,2,3,4}(u_1, u_2, u_3, u_4) = {} & c_{1,2}(u_1, u_2)c_{2,3}(u_2, u_3)c_{3,4}(u_3, u_4) \\ & c_{1,3;2}(F_{1|2}(u_1|u_2), F_{3|2}(u_3|u_2)) \\ & c_{2,4;3}(F_{2|3}(u_2|u_3), F_{4|3}(u_4|u_3)) \\ & c_{1,4;2,3}(F_{1|2,3}(u_1|u_3, u_3), F_{4|2,3}(u_4|u_2, u_3)) \end{aligned}$$

Now each of the bivariate copulas, i.e., each $c_{i,j;\mathbf{v}}$ are chosen by us, and therefore does not require further calculation to be evaluated. In contrast, the conditional marginal distributions have to be computed, and we now explain how this is done. We note that for general regular vines, there is a simple recursive method to calculate the required conditional densities, see Section 2.4 of Dissmann *et al.* (2013).

For the lowest tree, the marginals are uniform, and we do not need to deal with them. For the second tree, use

$$F_{i|j}(u_i|u_j) = \frac{\partial}{\partial u_j} C_{i,j}(u_i, u_j)$$

as discussed above. For the third, and here highest tree, we need to compute $F_{1|2,3}$ and $F_{4|2,3}$. Since we assume the vine copula has a density, $F_{1|2,3}$ is the cumulative distribution function of the density

$$f_{1|23}(u_1|u_2, u_3) = \frac{c_{1,2,3}(u_1, u_2, u_3)}{c_{2,3}(u_2, u_3)}. \tag{9}$$

Now, $(U_1, U_2, U_3)$ is also generated from a vine, and this vine can be found by removing everything that has to do with the other variables. Here, this sub-vine results in exactly the three-dimensional vine we used in earlier examples, see Figure 4. Therefore, we know that the joint distribution of $(U_1, U_2, U_3)$ has joint density

$$c_{1,2,3}(u_1, u_2, u_3) = c_{1,2}(u_1, u_2)c_{2,3}(u_2, u_3)c_{1,3;2}(F_{1|2}(u_1|u_2), F_{3|2}(u_3|u_2)).$$

Inserting this into (9) gives

$$\begin{aligned} f_{1|23}(u_1|u_2,u_3) &= \frac{c_{1,2,3}(u_1,u_2,u_3)}{c_{2,3}(u_2,u_3)} \\ &= c_{1,2}(u_1,u_2)c_{1,3;2}(F_{1|2}(u_1|u_2),F_{3|2}(u_3|u_2)) \\ &= c_{1,2}(u_1,u_2)c_{1,3;2}\left(\frac{\partial}{\partial u_2}C_{1,2}(u_1,u_2),\frac{\partial}{\partial u_2}C_{2,3}(u_2,u_3)\right) \end{aligned}$$

Recalling that we wish to identify not the density $f_{1|23}$ but instead the cumulative distribution function $F_{1|23}$, we integrate with respect to $u_1$. Now for $u_1 < 0$ or $u_1 > 1$ we have $c_{1,2}(u_1,u_2) = 0$, and therefore we start integrating at 0, and get, for $u_1 \leq 1$, that

$$F_{1|23}(u_1|u_2,u_3) = \int_0^{u_1} c_{1,2}(x_1,u_2)c_{1,3;2}\left(\frac{\partial}{\partial u_2}C_{1,2}(x_1,u_2),\frac{\partial}{\partial u_2}C_{2,3}(u_2,u_3)\right)\,dx_1.$$

Using the substitution

$$y = \frac{\partial}{\partial u_2}C_{1,2}(x_1,u_2),$$

which has derivative

$$\frac{d}{dx_1}y = \frac{\partial^2}{\partial x_1\partial u_2}C_{1,2}(x_1,u_2) = c_{1,2}(x_1,u_2),$$

integration with substitution gives

$$\begin{aligned} F_{1|23}(u_1|u_2,u_3) &= \int_0^{y(u_1)} c_{1,3;2}\left(y,\frac{\partial}{\partial u_2}C_{2,3}(u_2,u_3)\right)\,dy \qquad (10) \\ &= \int_0^{y(u_1)} \frac{\partial^2}{\partial u_1\partial u_2}C_{1,3;2}\left(y,\frac{\partial}{\partial u_2}C_{2,3}(u_2,u_3)\right)\,dy \\ &= D_2C_{1,3;2}\left(y(u_1),D_2C_{2,3}(u_2,u_3)\right) \\ &= D_2C_{1,3;2}\left(D_2C_{1,2}(u_1,u_2),D_2C_{2,3}(u_2,u_3)\right). \end{aligned}$$

where we, to avoid notational ambiguity, use the notation $D_iH(x_1,x_2) = (\partial/\partial x_i)H(x_1,x_2)$. By a similar argument, we identify $F_{4|23}$. We have that

$$f_{4|23}(u_4|u_2,u_3) = \frac{c_{2,3,4}(u_2,u_3,u_4)}{c_{2,3}(u_2,u_3)},$$

where the density of the sub-vine $(U_2,U_3,U_4)$ is deduced by the previously described general technique, giving

$$c_{2,3,4}(u_2,u_3,u_4) = c_{2,3}(u_2,u_3)c_{3,4}(u_3,u_4)c_{2,4;3}(F_{2|3}(u_2|u_3),F_{4|3}(u_4|u_3))$$

and therefore

$$f_{4|23}(u_4|u_2,u_3) = c_{3,4}(u_3,u_4)c_{2,4;3}(D_2C_{2,3}(u_2,u_3),D_1C_{3,4}(u_3,u_4)).$$

It then follows that

$$F_{4|2,3}(u_4|u_2,u_3) = D_1C_{2,4;3}(D_2C_{2,3}(u_2,u_3),D_1C_{3,4}(u_3,u_4)). \qquad (11)$$

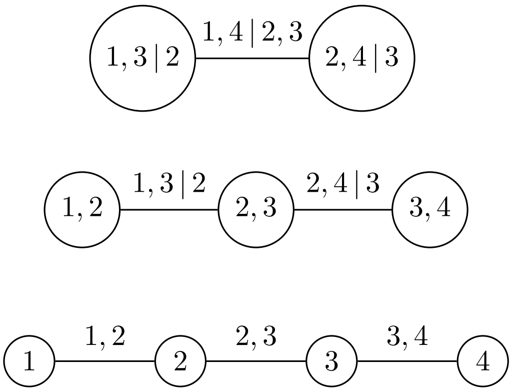


Figure 10: A four-dimensional regular vine.

Combining the above derivations gives a complete expression for the density of the vine in Figure 10. A notable feature is that numerical integration is avoided, at least when the densities and cumulative distribution functions of the bivariate copulas chosen by the user can be evaluated without numerical integration, which is usually the case for commonly used bivariate copulas. In cases where numerical integration is required, only bivariate numerical integration is needed, which is considerably less complex than general high-dimensional integration.

## A.5. How to simulate from a vine

Since the three-dimensional case considered earlier is too simple to easily see the general pattern of how to simulate from a general multivariate regular vine, we here consider the four-dimensional vine of Figure 10. The four-dimensional case is sufficiently complex for the general case to be within reach after having studied it.

### *Simulation from a general p-dimensional copula*

We start by providing a general algorithm for simulating from an arbitrary $p$-dimensional copula. This method, while general, will in high dimensions often be numerically infeasible, as there are no closed form expressions for the quantities required for applying the algorithm and numerical approximations have to be employed. In contrast, we will see that simulating from vines is computationally simpler, since high-dimensional numerical integration is not required in most cases. Vine simulation, illustrated via a four-dimensional example, will be explained in the next section.

The general simulation method we now present extends the bivariate and trivariate examples given above, and continues to use the multivariate quantile transformation. For $p$ variables this transform takes the following form (Rüschendorf 2009, Section 3). We want to simulate from a $p$-dimensional copula CDF $C$. We simulate $V_1, V_2, \ldots, V_p$ which are independent and uniform on $[0, 1]$. Then we let $U_1 = V_1$ (the marginals are already uniform), and recursively define

$$U_j = F^{-1}_{j|1,2,\ldots,j-1}(V_j|U_1, U_2, \ldots, U_{j-1}),$$

where $F^{-1}_{j|1,2,\ldots,j-1}$ is the generalized inverse of the conditional distribution function

$$F_{j|1,2,\ldots,j-1}(u_j|u_1,u_2,\ldots,u_{j-1}).$$

For simplicity, we will assume that $C$ has a density. Let $C_{1,2,\ldots,j}$ denote the distribution of $(U_1,U_2,\ldots,U_j)$, given by $C_{1,2,\ldots,j}(u_1,u_2,\ldots,u_j) = C(u_1,u_2,\ldots,u_j,1,1,\ldots,1)$, and let $c_{1,2,\ldots,j}$ denote its density. For simplicity, we assume that the density $c_{1,2,\ldots,j}$ is strictly positive for all inner points in the unit cube $[0,1]^j$. Recall that the density of a subset of the variables, such as the density $c_{1,2,\ldots,j}$ can be found by $c_{1,2,\ldots,j}(u_1,u_2,\ldots,u_j) = \partial^j C(u_1,u_2,\ldots,u_j,1,1,\ldots,1)/(\partial u_1\cdots\partial u_j)$. We have

$$\begin{aligned}
&F_{j|1,2,\ldots,j-1}(u_j|u_1,u_2,\ldots,u_{j-1})\\
&\qquad= \int_0^{u_j} \frac{c_{1,2,\ldots,j}(u_1,\ldots,x_j)}{c_{1,2,\ldots,j-1}(u_1,\ldots,u_{j-1})}\,dx_j = \frac{\int_0^{u_j} c_{1,2,\ldots,j}(u_1,\ldots,x_j)\,dx_j}{c_{1,2,\ldots,j-1}(u_1,\ldots,u_{j-1})}\\
&\qquad= \frac{\partial^{j-1}}{\partial u_1\partial u_2\cdots\partial u_{j-1}} C_{1,2,\ldots,j}(u_1,u_2,\ldots,u_j)\,(c_{1,2,\ldots,j-1}(u_1,\ldots,u_{j-1}))^{-1}
\end{aligned}$$

Notice that if, say, only the density $c$ of $C$ is known, the computation of $F_{j|1,2,\ldots,j-1}$ requires numerical integration routines in order to approximate the integral $\int_0^{u_j} c_{1,2,\ldots,j}(u_1,\ldots,x_j)\,dx_j$. The bivariate and trivariate simulation methods also follow the above pattern, and we have shown earlier that they work as intended. We may therefore conclude that the general method is valid using a proof by induction: Supposing this works for generating $(U_1,U_2,\ldots,U_{j-1})$, we prove that it also works for $(U_1,U_2,\ldots,U_j)$. By the induction hypothesis, $(U_1,U_2,\ldots,U_{j-1})$ is already generated as required. Then we generate

$$U_j = h^{-1}_{U_1,U_2,\ldots,U_{j-1}}(V_j),$$

where $h^{-1}_{u_1,u_2,\ldots,u_{j-1}}(u_j)$ is the (generalized) inverse function of $F_{j|1,2,\ldots,j-1}(u_j|u_1,u_2,\ldots,u_{j-1})$. Again we restrict attention to the cases where $h$ is in invertible in the regular sense. We have

$$\begin{aligned}
&P(U_1 \le u_1,\ldots,U_{j-1}\le u_{j-1},U_j\le u_j)\\
&\quad= \mathsf{E}(I\{U_1\le u_1,\ldots,U_{j-1}\le u_{j-1}\}\mathsf{E}[I\{h^{-1}_{U_1,U_2,\ldots,U_{j-1}}(V_j)\le u_j\}|U_1,\ldots,U_{j-1}])\\
&\quad= \mathsf{E}(I\{U_1\le u_1,\ldots,U_{j-1}\le u_{j-1}\}\mathsf{E}[I\{V_j\le h_{U_1,U_2,\ldots,U_{j-1}}(u_j)\}|U_1,\ldots,U_{j-1}])\\
&\quad= \mathsf{E}I\{U_1\le u_1,\ldots,U_{j-1}\le u_{j-1}\}h_{U_1,U_2,\ldots,U_{j-1}}(u_j)\\
&\quad= \mathsf{E}I\{U_1\le u_1,\ldots,U_{j-1}\le u_{j-1}\}F_{1,\ldots,j-1}(u_j|U_1,U_2,\ldots,U_{j-1})\\
&\quad= \int_0^{u_1}\cdots\int_0^{u_{j-1}} c_{1,\ldots,j-1}(x_1,\ldots,x_{j-1})F_{1,\ldots,j-1}(u_j|x_1,x_2,\ldots,x_{j-1})\,dx_1\cdots dx_{j-1}\\
&\quad= \int_0^{u_1}\cdots\int_0^{u_{j-1}} \frac{\partial^{j-1}}{\partial u_1\partial u_2\cdots\partial u_{j-1}} C_{1,2,\ldots,j}(x_1,x_2,\ldots,x_{j-1},u_j)\,dx_1\cdots dx_{j-1}\\
&\quad= C_{1,2,\ldots,j}(u_1,u_2,\ldots,u_j).
\end{aligned}$$

as required.

### *Simulating from a four-dimensional vine*

Let us now see how to apply this general technique to the four-dimensional vine copula distribution represented in Figure 10. Again, simulation can be performed without needing numerical integration. The general simulation approach from the multivariate quantile transform always simulates from $F_{j|1,2,\ldots,j-1}$ with $j$ starting at 1 and increasing up to $p$. This will always work, but for vines, we may simulate directly from the bivariate conditional distributions specified in the vine. Simulating from a bivariate conditional distribution will amount to simulate from two conditional distributions that are connected via bivariate copulas. This will in total lead to the same steps as the multivariate quantile transform, but has the advantage of having computations that are simpler to follow, as we follow the structure of the vine. A general simulation algorithm for regular vines is given in Algorithm 2.2 of Dissmann *et al.* (2013).

The main insight we need is that we have direct knowledge of certain conditional distributions from how the vine distribution is specified. We have easy access to the following conditional (and unconditional) distributions

$$\begin{aligned}&F_{1,4|2,3},\\&F_{1,3|2}, F_{2,4|3},\\&F_{1,2}, F_{2,3}, F_{3,4}.\end{aligned}$$

These distributions are all bivariate, and, as we have seen from constructing the joint density of $(U_1, U_2, U_3, U_4)$, can be joined to produce the full joint distribution of $(U_1, U_2, U_3, U_4)$.

We already know how to simulate from bivariate distributions. Let us see how this can be extended to simulating from bivariate conditional distributions. Suppose therefore, that we have simulated $(U_2, U_3)$ in such a way that it has the required bivariate distribution, i.e., it has the cumulative distribution function $C_{2,3}(u_2, u_3) = C(1, u_2, u_3, 1)$. We may do this directly using previously described techniques, since the copula and marginals of $U_2, U_3$ are known and directly specified.

To simulate the remaining coordinates $U_1, U_4$ we start by simulating from the conditional distribution of $U_1, U_4$ when conditioning on $U_2, U_3$, whose conditional CDF is denoted by $F_{1,4|2,3}$. By the simplifying assumption, the copula $C_{1,4;2,3}(u_1, u_4; u_2, u_3)$ of $F_{1,4|2,3}$ does not depend on $u_2, u_3$, and we therefore write $C_{1,4;2,3}(u_1, u_4; u_2, u_3) = C_{1,4;2,3}(u_1, u_4)$. Due to the simplifying assumption, $C_{1,4;2,3}$ is further a bivariate copula that we have chosen, which connects $U_1, U_4$ when conditioning on $U_2, U_3$. This implies that

$$F_{1,4|2,3}(u_1, u_3|u_2, u_3) = C_{1,4;2,3}(F_{1|2,3}(u_1|u_2, u_3), F_{4|2,3}(u_4|u_2, u_3)).$$

How should we simulate from $F_{1,4|2,3}$? We will show that if $u_2$ and $u_3$ are fixed to the already simulated $U_2$ and $U_3$ respectively, we may treat $F_{1,4|2,3}$ as if it is a standard (non-conditional) distribution, and use already described techniques to simulate from this bivariate distribution. The resulting variables will be valid simulations of the remaining $U_1, U_4$.

Since $F_{1,4|2,3}$ is a bivariate distribution with non-uniform marginals, we will as before split this simulation into two steps. Firstly, we simulate $W_1, W_4$ from its copula, which is $C_{1,4;2,3}$. By the simplifying assumption, this copula is a standard bivariate copula which does not depend on variables simulated earlier. We will then transform $W_1, W_4$ using univariate quantile transforms so that they have distributions $F_{1|2,3}$ and $F_{4|2,3}$ respectively.

Let us first simulate $W_1, W_4$ from the bivariate copula $C_{1,4;2,3}$. How this is done has been explained earlier: We simulate independent $V_1, V_4$ from $U[0, 1]$. Then we set $W_4 = V_4$ and

$$W_1 = h_{W_4}^{(1,4;2,3)}(V_1),$$

where

$$h_{w_4}^{(1,4;2,3)}(v_1) = D_1 C_{1,4;2,3}(v_1, w_4).$$

Again note that due to the simplifying assumption, this step does not depend on the already simulated $U_2, U_3$.

We next transform $W_1, W_4$ so that they are $F_{1|2,3}$ and $F_{3|2,3}$ distributed respectively. An important point here is that this is where dependence from $U_2$ and $U_3$ is introduced. We again use the univariate quantile transform and set

$$U_1 = F_{1|2,3}^{-1}(W_1|U_2, U_3), \qquad U_4 = F_{4|2,3}^{-1}(W_4|U_2, U_3),$$

where $F_{1|2,3}$ is given in (10) (p. 40) and $F_{4|2,3}$ is given in (11) (p. 40), considering the already simulated $U_2, U_3$ as fixed. It is not immediately apparent that $U_1, U_4$ have uniform marginals. This can be shown directly, but we will now show the more general fact that $(U_1, U_2, U_3, U_4)^\top \sim C$, and since all marginals in $C$ are uniform, this also implies that $U_1, U_4$ have uniform marginals.

Let us for completeness (and since we do not know an elementary reference for this fact) show the formal validity of this simulation method. That is, let the variables generated in this fashion be denoted $U_1, U_2, U_3, U_4$. We will show that the joint CDF of these variables is $C$. Again, all inverse functions are assumed to be traditional inverse functions. Since $(W_1, W_4)^\top \sim C_{1,4;2,3}$ is independent to $(U_2, U_3)$ we have

$$\begin{aligned}
&P(U_1 \le u_1, U_2 \le u_2, U_3 \le u_3, U_4 \le u_4) \\
&\quad = \mathsf{E} I\{U_2 \le u_2, U_3 \le u_3\} \mathsf{E}[I\{F_{1|2,3}^{-1}(W_1|U_2, U_3) \le u_1, F_{4|2,3}^{-1}(W_4|U_2, U_3) \le u_4\}|(U_2, U_3)] \\
&\quad = \mathsf{E} I\{U_2 \le u_2, U_3 \le u_3\} \mathsf{E}[I\{W_1 \le F_{1|2,3}(u_1|U_2, U_3), W_4 \le F_{4|2,3}(u_4|U_2, U_3)\}|(U_2, U_3)] \\
&\quad \overset{(a)}{=} \mathsf{E} I\{U_2 \le u_2, U_3 \le u_3\} \mathsf{E}_{W_1,W_2}[I\{W_1 \le F_{1|2,3}(u_1|U_2, U_3), W_4 \le F_{4|2,3}(u_4|U_2, U_3)\}] \\
&\quad = \mathsf{E} I\{U_2 \le u_2, U_3 \le u_3\} P_{W_1,W_2}\Big(W_1 \le F_{1|2,3}(u_1|U_2, U_3), W_4 \le F_{4|2,3}(u_4|U_2, U_3)\Big) \\
&\quad \overset{(b)}{=} \mathsf{E} I\{U_2 \le u_2, U_3 \le u_3\} C_{1,4;2,3}(F_{1|2,3}(u_1|U_2, U_3), F_{4|2,3}(u_4|U_2, U_3)) \\
&\quad = \int_0^{u_2} \int_0^{u_3} c_{2,3}(x_2, x_3) C_{1,4;2,3}(F_{1|2,3}(u_1|x_2, x_3), F_{4|2,3}(u_4|x_2, x_3)) \, dx_2 dx_3
\end{aligned}$$

Explanations: *(a)* Use that $(W_1, W_4)$ is independent to $(U_2, U_3)$. $\mathsf{E}_{W_1,W_2}$ and $P_{W_1,W_2}$ means that we integrate only over $W_1, W_2$. *(b)* Use that $(W_1, W_4)^\top \sim C_{1,4;2,3}$.

Since

$$C_{1,4;2,3}(F_{1|2,3}(u_1|x_2, x_3), F_{4|2,3}(u_4|x_2, x_3)) = F_{1,4|2,3}(u_1, u_4|x_2, x_3),$$

where $F_{1,4|2,3}$ is the conditional distribution of $(U_1, U_4)$ conditional on $(U_2, U_3)$, we have that

$$\begin{aligned}
C_{1,4;2,3}(F_{1|2,3}(u_1|x_2, x_3), F_{4|2,3}(u_4|x_2, x_3)) &= \int_0^{u_1} \int_0^{u_4} c_{1,4|2,3}(x_1, x_4|x_2, x_3) \, dx_1 x_4 \\
&= \int_0^{u_1} \int_0^{u_4} \frac{c_{1,2,3,4}(x_1, x_2, x_3, x_4)}{c_{2,3}(x_2, x_3)} \, dx_1 x_4.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&P(U_1 \le u_1, U_2 \le u_2, U_3 \le u_3, U_4 \le u_4) \\
&\quad = \int_0^{u_2} \int_0^{u_3} c_{2,3}(x_2, x_3) C_{1,4;2,3}(F_{1|2,3}(u_1|x_2, x_3), F_{4|2,3}(u_4|x_2, x_3))\, dx_2 x_3 \\
&\quad = \int_0^{u_2} \int_0^{u_3} c_{2,3}(x_2, x_3) \int_0^{u_1} \int_0^{u_4} \frac{c_{1,2,3,4}(x_1, x_2, x_3, x_4)}{c_{2,3}(x_2, x_3)}\, dx_1 x_4 dx_2 x_3 \\
&\quad = \int_0^{u_1} \int_0^{u_2} \int_0^{u_3} \int_0^{u_4} c_{1,2,3,4}(x_1, x_2, x_3, x_4) dx_1 dx_2 dx_3 dx_4 \\
&\quad = C(u_1, u_2, u_3, u_4),
\end{aligned}$$

showing the validity of the simulation.

## Affiliation:

Steffen Grønneberg
Department of Economics
BI Norwegian Business School
Oslo, Norway 0484
E-mail: steffeng@gmail.com

Njål Foldnes
Department of Economics
BI Norwegian Business School
Stavanger, Norway 4014
E-mail: njal.foldnes@gmail.com

Katerina M. Marcoulides
Department of Psychology
University of Minnesota
Elliott Hall 75 East River Rd.
Minneapolis, MN 55455, United States of America
E-mail: kmarcoul@umn.edu

*Journal of Statistical Software* https://www.jstatsoft.org/
published by the Foundation for Open Access Statistics https://www.foastat.org/
April 2022, Volume 102, Issue 3 *Submitted:* 2020-07-10
doi:10.18637/jss.v102.i03 *Accepted:* 2021-09-05