Generating random correlation matrices with fixed values: An application to the evaluation of multivariate surrogate endpoints

Peer-reviewed author version

FLOREZ POVEDA, Alvaro; ALONSO ABAD, Ariel; MOLENBERGHS, Geert & VAN DER ELST, Wim (2020) Generating random correlation matrices with fixed values: An application to the evaluation of multivariate surrogate endpoints. In: COMPUTATIONAL STATISTICS & DATA ANALYSIS, 142 (Art N° 106834).

DOI: 10.1016/j.csda.2019.106834
Handle: http://hdl.handle.net/1942/30477

Generating random correlation matrices with fixed values: An application to the evaluation of multivariate surrogate endpoints

Alvaro Jóse Flórez[a,*], Ariel Alonso Abad[b], Geert Molenberghs[a,b], Wim Van Der Elst[c]

[a] *I-BioStat, Universiteit Hasselt, B-3590 Diepenbeek, Belgium*
[b] *I-BioStat, KU Leuven, B-3000 Leuven, Belgium*
[c] *Janssen Pharmaceuticals, B-2340 Beerse, Belgium*

Abstract

When assessing surrogate endpoints in clinical studies under a causal-inference framework, a simulation-based sensitivity analysis is required, so as to sample the unidentifiable parameters across plausible values. To be precise, correlation matrices need to be sampled with only some of their entries identified from the data, known as the matrix completion problem. The positive-definiteness constraints are cumbersome functions involving all matrix entries, making this a challenging task. Some existing algorithms rely on sampling and then rejecting invalid solutions. A very efficient algorithm is built on previous work to generate large correlation matrices with some a prior fixed elements. The proposed methodology is applied to tackle a difficult problem in the surrogate marker field, namely, the evaluation of multivariate, potentially high-dimensional, surrogate endpoints. Whereas existing methods are limited to very low-dimensional surrogates, the new proposal is stable, fast, shows good properties, and is implemented in a user-friendly and freely available R package.

Keywords: Multiple surrogate evaluation; Partial correlation; Positive-definite matrix; Random correlation matrices; Simulation-based sensitivity analysis.

*Corresponding author: Alvaro J. Flórez, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium. Tel: +32 11268201.
Email address: alvaro.florez@uhasselt.be (Alvaro Jóse Flórez)

1. Introduction

In the causal-inference framework, one frequently fits models with an only partially identifiable set of parameters $\boldsymbol{\theta}$, i.e., there is a subset of $\boldsymbol{\theta}$ that cannot be estimated from the data. A possible solution to this problem is to impose untestable restrictions, e.g., based on expert knowledge, for the unidentifiable parameters to estimate the model. Alternatively, one can conduct a sensitivity analysis to assess how the fitted model and conclusions based there upon change as the unidentifiable parameters vary across plausible values. The latter option is taken by [1] and [2] to evaluate univariate and multivariate surrogate endpoints in a causal-inference framework. Surrogacy is then evaluated using the so-called individual causal association (ICA), i.e., the association between the individual causal treatment effects on the surrogate and true endpoints. The ICA is a function of a partially identifiable correlation matrix ($\boldsymbol{R}$). Their approach rests upon computing the ICA across a set of randomly generated correlation matrices, taken to mean sampling from the collection of all symmetric positive semi-definite matrices (PSD) of a given dimension, and with unit diagonal, but, importantly, while keeping the estimable values fixed. This so-called matrix completion problem is non-trivial.

The PSD constraint involves all values of a correlation matrix, and therefore, its random generation is very challenging. However, several algorithms have been presented. [3] proposes a method based on a transformation of partial correlations, later extended by [4]. The parameterization in terms of partial correlations and its application on the completion problem are also presented by [5] and [6], respectively. [7] introduced an algorithm using the hyperspherical parametrization (HP) of the Cholesky factor. More alternatives can be found in [8, 9, 10], among others. Although many of these approaches would target PSD matrices, we do prefer a positive-definite (PD) constraint because of the operations we need to perform on the so-resulting matrices, involving inversion of these as a whole and sub-matrices thereof.

In this paper, we build on previous work and evaluate through simulations

various algorithms to generate random correlation matrices with fixed entries in the multiple surrogaty assessment. Joe's algorithm is generalized by conveniently rearranging the fixed elements of the correlation matrix and leads to excellent performance, in particular also in terms of speed and the ability of handle high-dimensional matrices. On the other hand, the adaptation of the method proposed by [7] is cumbersome. Another alternative is to simulate a pseudo-correlation matrix and to find the nearest correlation matrix, according to some metric, keeping the identifiable values fixed. Some of these adjustments for non-positive-definiteness can be found in [11] and [12]. Furthermore, the methodology is applied to solve a difficult problem in the surrogate marker field, namely, the evaluation of multivariate surrogate endpoints, as well as in a different, high-dimensional context.

The structure of the manuscript is as follows. Section 2 presents the methodology for assessing multiple surrogates. In Section 3, various algorithms to generate unrestricted random correlation matrices are introduced. Section 4 describes the algorithms to generate random correlation matrices with some of their values fixed. A simulation study to compare the methods is executed in Section 5. A motivating experiment on mice (the transPAT study) is presented and analyzed in Section 6. Section 7 is reserved for final remarks.

2. Assessing a multivariate surrogate

In clinical trials, a true endpoint is defined as the most credible indicator of drug response. However, its measurement might be costly, difficult or requiring long follow-up time. Therefore, finding a less complex valid "substitute", termed as surrogate, of the true endpoint is very convenient [13]. In the last decades, several statistical methodologies to evaluate surrogate endpoints have been proposed, most of them within the causal-inference and meta-analytic paradigms. In this paper, we focus on the former. Details on surrogacy evaluation can be found in [13], [14], [15], among others.

We consider a single-trial setting: the data consist of measurements of

a univariate true endpoint T and a p-dimensional surrogate endpoint $\boldsymbol{S} = (S_1, \ldots, S_p)'$ for N patients. Moreover, only two treatments are under evaluation ($Z = 0/1$) in a parallel study design. Rubin's model for causal inference [16] assumes that each patient has two potential outcomes for T: an outcome T_0 that would be observed under the control treatment ($Z = 0$), and an outcome T_1 that would be observed under the experimental treatment ($Z = 1$). Furthermore, using obvious notation, let us now consider the $2\,(p+1)$ dimensional vector of potential outcomes $\boldsymbol{Y} = (T_0, T_1, S_{10}, S_{11}, S_{20}, S_{21}, \ldots, S_{p0}, S_{p1})'$ and the corresponding vector of individual causal treatment effects $\boldsymbol{\Delta} = (\Delta T, \boldsymbol{\Delta S}')'$, where $\Delta T = T_1 - T_0$ and $\boldsymbol{\Delta S} = (\Delta S_1, \Delta S_2, \ldots ..., \Delta S_p)'$ with $\Delta S_k = S_{k1} - S_{k0}$. The so-called fundamental problem of causal inference states that only one of the potential outcomes associated with the true and surrogate endpoints are observed in practice. Therefore, $\boldsymbol{\Delta}$ cannot be estimated from the data [17]. Note that, to avoid clutter, no subindex has been used to denote the patient.

Based on $\boldsymbol{\Delta}$, one can define the expected or average causal treatment effects in the population of interest as $E(\boldsymbol{\Delta}) = (\beta, \boldsymbol{\alpha}')'$, where $\beta = E(\Delta T)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)'$ with $\alpha_k = E(\Delta S_k)$.

[18] provided three identifiability conditions under which it is possible to obtain consistent estimators of the expected causal treatment effects. If Y denotes the response of interest and Y_z the potential outcome associated with $Z = z$ then the three identifiability conditions are: 1) Consistency: If $Z = z$ for a given subject then $Y_z = Y$ for that subject, 2) Conditional exchangeability: This condition essentially states that there are no unmeasured confounders given data on baseline covariates L, that is, $Y_z \perp Z | L = l$ for each possible value z of Z and l of L and 3) Positivity: If $f_L(l) \neq 0$ then $f_{Z|L}(z|l) > 0$. It can be easily shown that in randomized clinical trials, all condition hold, and the expected causal treatment effects can be estimated as $\beta = E(T|Z = 1) - E(T|Z = 0)$ and $\alpha_k = E(S_k|Z = 1) - E(S_k|Z = 0)$, where the conditional expectations are estimated using the observed means in the control and treated groups, respectively. The metric of surrogacy proposed by [2], and used in the following sections, is based only on the individual causal treatment effects and it is valid if consistency

holds, i.e., it could also be applied to observational data.

In the surrogacy evaluation context, one is interested in the distribution of the vector of potential outcomes $\boldsymbol{Y}$. Further, it will be further assumed that $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \boldsymbol{DRD})$, where $\boldsymbol{\mu} = (\mu_{T_0}, \mu_{T_1}, \mu_{S_{10}}, \mu_{S_{11}}, \ldots, \mu_{S_{p0}}, \mu_{S_{p1}})'$, $\boldsymbol{D}$ is a diagonal matrix with $(\sigma_{T_0}, \sigma_{T_1}, \sigma_{S_{10}}, \sigma_{S_{11}}, \ldots, \sigma_{S_{p0}}, \sigma_{S_{p1}})$ along the diagonal and

$$\boldsymbol{R} = \begin{pmatrix} 1 & \rho_{T_0T_1} & \boldsymbol{\rho_{T_0S_{10}}} & \rho_{T_0S_{11}} & \boldsymbol{\rho_{T_0S_{20}}} & \rho_{T_0S_{21}} & \cdots & \boldsymbol{\rho_{T_0S_{p0}}} & \rho_{T_0S_{p1}} \\ & 1 & \rho_{T_1S_{10}} & \boldsymbol{\rho_{T_1S_{11}}} & \rho_{T_1S_{20}} & \boldsymbol{\rho_{T_1S_{21}}} & \cdots & \rho_{T_1S_{p0}} & \boldsymbol{\rho_{T_1S_{p1}}} \\ & & 1 & \rho_{S_{10}S_{11}} & \boldsymbol{\rho_{S_{10}S_{20}}} & \rho_{S_{10}S_{21}} & \cdots & \boldsymbol{\rho_{S_{10}S_{p0}}} & \rho_{S_{10}S_{p1}} \\ & & & 1 & \rho_{S_{11}S_{20}} & \boldsymbol{\rho_{S_{11}S_{21}}} & \cdots & \rho_{S_{11}S_{p0}} & \boldsymbol{\rho_{S_{11}S_{p1}}} \\ & & & & 1 & \rho_{S_{20}S_{21}} & \cdots & \boldsymbol{\rho_{S_{20}S_{p0}}} & \rho_{S_{20}S_{p1}} \\ & & & & & 1 & \cdots & \rho_{S_{21}S_{p0}} & \boldsymbol{\rho_{S_{21}S_{p1}}} \\ & & & & & & \ddots & \vdots & \vdots \\ & & & & & & & 1 & \rho_{S_{p0}S_{p1}} \\ & & & & & & & & 1 \end{pmatrix}, \quad (1)$$

where bold symbols denote identifiable entries in $\boldsymbol{R}$. Note that (1) has an specific structure in which the unidentifiable parameters are located in the $(2k-1)$-diagonals, with $k = 1, \ldots, p$ of $\boldsymbol{R}$, i.e., $\rho_{i,j}$ is unestimable from the data when one of the pairs (i, j) is an even integer and the other is odd.

Under the previous assumptions, one has that $\boldsymbol{\Delta} \sim N(\boldsymbol{\mu_\Delta}, \boldsymbol{\Sigma_\Delta})$, with $\boldsymbol{\mu_\Delta} = (\beta, \boldsymbol{\alpha})'$ and

$$\boldsymbol{\Sigma_\Delta} = \begin{pmatrix} \sigma_{\Delta T} & \boldsymbol{\Sigma}'_{\boldsymbol{\Delta S}\Delta T} \\ \boldsymbol{\Sigma_{\Delta S}}_{\Delta T} & \boldsymbol{\Sigma_{\Delta S}} \end{pmatrix},$$

where $\sigma_{\Delta T} = \sigma^2_{T_0} + \sigma^2_{T_1} - 2\rho_{T_0T_1}\sqrt{\sigma^2_{T_0}\sigma^2_{T_1}}$ is the variance of ΔT; $\boldsymbol{\Sigma_{\Delta S}}_{\Delta T}$ is a p-dimensional vector of covariances between ΔT and $\boldsymbol{\Delta S}$; and $\boldsymbol{\Sigma_{\Delta S}}$ is the $(p \times p)$ variance-covariance matrix of $\boldsymbol{\Delta S}$.

2.1. Individual causal association based on a multivariate surrogate

In the univariate setting ($p = 1$), [1] defined the ICA as the Pearson correlation coefficient between ΔT and ΔS:

$$\rho_\Delta = \frac{\sigma_{T_0}\sigma_{S_0}\rho_{T_0S_0} + \sigma_{T_1}\sigma_{S_1}\rho_{T_1S_1} - \sigma_{T_1}\sigma_{S_0}\rho_{T_1S_0} - \sigma_{T_0}\sigma_{S_1}\rho_{T_0S_1}}{\sqrt{\left(\sigma^2_{T_0} + \sigma^2_{T_1} - 2\sigma_{T_0}\sigma_{T_1}\rho_{T_0T_1}\right)\left(\sigma^2_{S_0} + \sigma^2_{S_1} - 2\sigma_{S_0}\sigma_{S_1}\rho_{S_0S_1}\right)}}, \quad (2)$$

where ρ_{XY} is the correlation between the potential outcomes X and Y. It quantifies how accurate the prediction is of the causal treatment effect on the true endpoint using the causal treatment effect on the surrogate, for a given patient. Note that, the ICA is not simply the treatment-corrected correlation between S and T, but it is based on the individual causal treatment effects concept and it has a direct causal interpretation [1, 19].

Only the variances and correlations $\rho_{T_0S_0}$, $\rho_{T_1S_1}$ are identifiable from the data. Therefore, the ICA cannot be identified without imposing untestable restrictions on the unidentifiable correlations. Hence, [1] proposed a simulation-based sensitivity analysis in which ρ_Δ is calculated across a set of plausible values for the inestimable elements in (2). To be precise, they considered all positive-definite matrices over a grid of values $G = (g_1, \ldots, g_k)$, with $-1 \leq g_i \leq 1$ for the unidentifiable correlations, and then calculated ρ_Δ for each of them. The so-obtained collection produces an insightful distribution that can be examined graphically, or summarized using appropriate point and/or interval summaries.

For the case of a multivariate surrogate, [2] proposed the so-called squared information coefficient of correlation (SICC; 20, 21) to quantify the ICA, i.e.,

$$R_H^2 = \frac{\boldsymbol{\Sigma}'_{\boldsymbol{\Delta S}\Delta T}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\Delta S}}\boldsymbol{\Sigma}_{\boldsymbol{\Delta S}\Delta T}}{\sigma_{\Delta T}}.$$

The R_H^2 ranges over the unit interval $[0, 1]$, and takes the value zero if and only ΔT and $\boldsymbol{\Delta S}$ are independent, while it takes the value one if and only if ΔT is perfectly linearly predictable from $\boldsymbol{\Delta S}$. As in the univariate case, R_H^2 is not identifiable from the data and a simulation-based sensitivity analysis is recommended. However, the grid-based approach becomes computationally too intensive, even infeasible, as the number of surrogate endpoints increases. The main issue is that with increasing dimensions, the space of positive-definite matrices is an ever smaller subset of the rectangle $[-1, +1]^k$ with k as the number of functionally different correlations involved. Therefore, this approach, rejection sampling, to be discussed in the next section, is limited to a small number of surrogates.

3. Generating random correlation matrices

3.1. Rejection sampling (RS) algorithm

In this method, $\boldsymbol{R}$ is constructed by drawing ρ_{ij} independently from a distribution on $(-1, 1)$, e.g., uniform. If the generated $\tilde{\boldsymbol{R}}$ matrix is not PD, it is rejected. This method performs well when the order of the correlation matrix is small. However, the rejection rate increases rapidly with d, making it time-consuming or even not possible. [22] derived an expression for the probability of the RS algorithm to generate a valid correlation matrix, and the rejection probability is practically equal to one for $d \geq 6$.

3.2. Gradual rejection sampling (GRS) algorithm

To decrease the rejection rate and computing time, the RS algorithm can be implemented in a gradual way using Sylvester's criterion, i.e., a matrix is PD when all the upper-left sub-matrices have positive determinants. Thus, to generate a $(d \times d)$ correlation matrix, we start by randomly sampling the upper-left (2×2) sub-matrix using the RS algorithm **(i.e., simulating ρ_{12})**. When the determinant is positive, the same procedure is used for the upper-left (3×3) sub-matrix **(i.e., sampling ρ_{13} and ρ_{24} keeping ρ_{12} fixed)**, and so on **until the** $(d \times d)$ **is completely generated (in the last step, $\rho_{1d}, \rho_{2d}, \ldots, \rho_{d-1,d}$ are simulated keeping the other correlation fixed)**. This approach improves the acceptance rate and speed of the RS algorithm. However, it is still limited in practice to $d \leq 10$ [2].

3.3. Algorithm based on partial correlations (PC)

[3] proposed to generate a PD random correlation matrix $\boldsymbol{R}$ progressively based on a parameterization in terms of the correlations $\rho_{i,i+1}$ for $i = 1, \ldots, d-1$ and the partial correlations $\rho_{ij|i+1,\ldots,j-1}$ for $j - i \geq 2$. The algorithm is based on the following equality:

$$\rho_{j,j+k} = \boldsymbol{r}_1'(j,k) \left\{\boldsymbol{R}_2(j,k)\right\}^{-1} \boldsymbol{r}_3(j,k) + \rho_{j,j+k|j+1,\ldots,j+k-1} D_{j,k}, \qquad (3)$$

where

$$\boldsymbol{R}[j:j+k] = \begin{pmatrix} 1 & \boldsymbol{r}_1'(j,k) & \rho_{j,j+k} \\ \boldsymbol{r}_1(j,k) & \boldsymbol{R}_2(j,k) & \boldsymbol{r}_3(j,k) \\ \rho_{j+k,j} & \boldsymbol{r}_3'(j,k) & 1 \end{pmatrix},$$

with $\boldsymbol{r}_1'(j,k) = (\rho_{j,j+1}, \ldots, \rho_{j,j+k-1})$, $\boldsymbol{r}_3'(j,k) = (\rho_{j+k,j+1}, \ldots, \rho_{j+k,j+k-1})$, $\boldsymbol{R}_2(j,k)$ being the middle $(k-1 \times k-1)$ matrix of $\boldsymbol{R}[j:j+k]$, and:

$$D_{j,k} = \sqrt{\{1 - \boldsymbol{r}_1'(j,k)\boldsymbol{R}_2(j,k)\boldsymbol{r}_1(j,k)\}\{1 - \boldsymbol{r}_3'(j,k)\boldsymbol{R}_2(j,k)\boldsymbol{r}_3(j,k)\}}.$$

Then, one can generate $\rho_{j,j+k|j+1,\ldots,j+k-1}$ $(1 \leq k \leq d-1)$ independently in the interval $(-1,1)$ and then use (3) to get $\rho_{i,i+k}$ for $2 \leq k \leq d-1$. Furthermore, the correlations $\rho_{j,j+1}$ are also independently generated in the interval $(-1,1)$. Therefore, a correlation matrix of size $(d \times d)$ is generated by sampling $\binom{d}{2}$ appropriately chosen partial correlations. [4] extended the method to allow computationally more efficient choices of $\binom{d}{2}$ partial correlations.

To obtain identical symmetric marginal densities of each $\rho_{i,j}$, [3] proposes to draw each $\rho_{j,j+1}$ $(j = 1, \ldots, d-1)$ from a Beta$\left(\frac{d}{2}, \frac{d}{2}\right)$ on $(-1,1)$, and each $\rho_{j,j+k|j+1,\ldots,j+k-1}$ from a Beta$\left\{1 + \frac{1}{2}(d-1-k), 1 + \frac{1}{2}(d-1-k)\right\}$ on $(-1,1)$.

3.4. Algorithm based on the hyperspherical parameterization of the Cholesky factor (HP)

[7] introduced a method based on the reparameterization of the Cholesky factor of $\boldsymbol{R}$ using hyperspherical coordinates. Following the Cholesky factorization, a PD correlation matrix $\boldsymbol{R}$ can be factorized by $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}'$, where $\boldsymbol{U} = (u_{ij})$ is a lower triangular matrix with $u_{11} = 1$, $u_{i1} = \cos\theta_{i,1}$, for $i = 1, \ldots, d$, and

$$u_{ij} = \begin{cases} \prod_{k=1}^{j-1} \sin\theta_{ik} & \text{for } i = j, \\ \cos\theta_{ij} \prod_{k=1}^{j-1} \sin\theta_{ik} & \text{for } 2 \leq j \leq i-1, \end{cases}$$

where $\theta_{ij}, i > j$, are angles restricted to $(0, \pi)$. Furthermore, the transformation from $\boldsymbol{R}$ to $\boldsymbol{\Theta} = (\theta_{ij})$ is one-to-one.

Then, a random correlation matrix $\boldsymbol{R}$ is generated by drawing values of the $j-$th columns of the lower-triangular matrix $\boldsymbol{\Theta}$ using the following distribution:

$$\theta_{ij} \sim g_j(\theta) \propto (\sin\theta)^{2k+n-j} I(0 < \theta < \pi), i = j+1, \ldots, d, \tag{4}$$

where $k \geq 0$; and later, computing the lower triangular matrix $\boldsymbol{U}$ using $\theta_{i,j}$ and constructing the correlation matrix $\boldsymbol{R} = \boldsymbol{U}\boldsymbol{U}'$.

For $k = 0$, each ρ_{ij} follows a Beta$\left(\frac{d}{2}, \frac{d}{2}\right)$ distribution on $(-1, 1)$, leading to the same distribution as the algorithm proposed by [3].

4. Generating correlation matrices with fixed values

With some correlations fixed, the same algorithms as presented in Section 3 can be implemented. The RS and gradual-RS algorithms can be applied directly by randomly drawing the non-fixed $\rho_{i,j}$ values independently from a uniform distribution on $(-1, 1)$ and rejecting the matrix if it is not PD. The latter is implemented by [2] to compute R^2_H. However, they concluded that this algorithm is computationally suitable to consider at most four surrogates, which corresponds to generate random correlations matrices of size 10. For a larger number of surrogates, it breaks down.

The implementation of the PC algorithm, or its extensions, for the matrix completion depends on the pattern of the fixed entries. Using setting (1), its application is complicated. However, we can rearrange $\boldsymbol{R}$ as follows:

$$\boldsymbol{R}' = \begin{pmatrix}
1 & \boldsymbol{\rho_{T_0 S_{10}}} & \boldsymbol{\rho_{T_0 S_{20}}} & \cdots & \boldsymbol{\rho_{T_0 S_{p0}}} & \rho_{T_0 T_1} & \rho_{T_0 S_{11}} & \rho_{T_0 S_{21}} & \cdots & \rho_{T_0 S_{p1}} \\
 & 1 & \boldsymbol{\rho_{S_{10} S_{20}}} & \cdots & \boldsymbol{\rho_{S_{10} S_{p0}}} & \rho_{S_{10} T_1} & \rho_{S_{10} S_{11}} & \rho_{S_{10} S_{21}} & \cdots & \rho_{S_{10} S_{p1}} \\
 & & 1 & \cdots & \boldsymbol{\rho_{S_{20} S_{p0}}} & \rho_{S_{20} T_1} & \rho_{S_{20} S_{11}} & \rho_{S_{20} S_{21}} & \cdots & \rho_{S_{20} S_{p1}} \\
 & & & \ddots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
 & & & & 1 & \rho_{S_{p0} T_1} & \rho_{S_{p0} S_{11}} & \rho_{S_{p0} S_{21}} & \cdots & \rho_{S_{p0} S_{p1}} \\
 & & & & & 1 & \boldsymbol{\rho_{T_1 S_{11}}} & \boldsymbol{\rho_{T_1 S_{21}}} & \cdots & \boldsymbol{\rho_{T_1 S_{p1}}} \\
 & & & & & & 1 & \boldsymbol{\rho_{S_{11} S_{21}}} & \cdots & \boldsymbol{\rho_{S_{11} S_{p1}}} \\
 & & & & & & & 1 & \cdots & \boldsymbol{\rho_{S_{21} S_{p1}}} \\
 & & & & & & & & \ddots & \vdots \\
 & & & & & & & & & 1
\end{pmatrix}. \tag{5}$$

Starting from (5), we can apply the PC algorithm by generating partial

correlations to complete one diagonal at a time starting from the diagonal closest to the middle.

The adaptation of the HP algorithm is not straightforward. Translating the constraints on $\rho_{i,j}$ to the angles $\theta_{i,j}$ is cumbersome and unpractical, even for small d. Another alternative is to generate a random correlation matrix ($\boldsymbol{R}^{(0)}$) using the HP algorithm, replacing the prior fixed values to yield $\boldsymbol{R}^{(1)}$. If the $\boldsymbol{R}^{(1)}$ is not PD, we can find the nearest PD correlation matrix using the weighted scaling or linear shrinking method [11].

4.1. Weighted scaling method

Here, we find the PD correlation matrix $\tilde{\boldsymbol{R}} = (\tilde{r}_{i,j})$ that is as near as possible to the pseudo-correlation matrix $\boldsymbol{R}^{(1)} = (r^{(1)}_{i,j})$ using the following criterion:

$$S = \sum_{i=1}^{d} \sum_{i=1}^{d} w_{ij} \left(r^{(1)}_{ij} - \tilde{r}_{ij} \right)^2, \tag{6}$$

where $w_{ij} = w_{ji}$ is the weight associated to correlation r_{ij}. Given that the PD constraint on $\tilde{\boldsymbol{R}}$ involves all $\tilde{r}_{ij}$ simultaneously, the minimization of (6) seems intractable. However, this problem is overcome by describing $\tilde{\boldsymbol{R}}$ in a geometric way (Section 3.4), and then, iteratively finding the angles θ_{ij} that minimize (6). By minimizing S as a function of θ_{ij}, the PD constraint on $\tilde{\boldsymbol{R}}$ is incorporated. Since S is differentiable with respect to θ_{ij}, we can apply iterative numerical algorithms, e.g., the steepest descent method. As initial values, we can use the θ_{ij} associated to $\boldsymbol{R}^{(1)}$.

To ensure that the fixed values do not move, or do so only very slightly, we set the corresponding weights to a very large value (e.g., 10^5) and, the weights for elements that are allowed to change are set equal to a small value, e.g, one.

4.2. Linear shrinking method

Here, the pseudo-correlation matrix $\boldsymbol{R}^{(1)}$ is shrunk towards an arbitrary correlation matrix $\boldsymbol{R}^{(0)}$ according to,

$$\tilde{\boldsymbol{R}} = \lambda \boldsymbol{R}^{(1)} + (1 - \lambda) \boldsymbol{R}^{(0)}, \tag{7}$$

where λ is the largest value in $[0, 1]$ which makes $\tilde{\boldsymbol{R}}$ PD. In our case, $\boldsymbol{R}^{(0)}$ contains zero values for all correlations, except for the fixed ones. Then, to find $\tilde{\boldsymbol{R}}^{(1)}$, we proceed as follows: (1) find λ in (7), (2) generate a random value $\tilde{\lambda}$ in $(0, \lambda)$, using an uniform distribution, for example, and (3) compute $\tilde{\boldsymbol{R}}$ according to (7) using $\tilde{\lambda}$.

5. Simulation study

To assess the performance of the previously introduced algorithms, a simulation study was carried out. The main objective of the simulation was to evaluate the computational feasibility of the different algorithms, i.e., to assess the time required to draw a random positive definite matrix. In addition, the generated random matrices were used to evaluate the validity of a putative multivariate surrogate endpoint $\boldsymbol{S} = (S_1, \ldots, S_p)'$ for a univariate true endpoint (T) using the ICA.

5.1. Settings

The identifiable correlations will be assumed equal, while different values will be considered: (a) for the number of surrogates, and therefore the matrix size and number of fixed entries, (b) and for the values of the fixed identifiable correlations:

- **Number of surrogates** (p)**:** one surrogate (i.e., a 4×4 matrix with two fixed values), three surrogates (i.e., a 8×8 matrix with 12 fixed values), five surrogates (i.e., a 12×12 matrix with 30 fixed values) and 10 surrogates (i.e., a 22×22 matrix with 110 fixed values).

- **Level of correlation** (ρ)**:** low ($\rho = 0.2$), moderate ($\rho = 0.5$), and high ($\rho = 0.8$) correlation.

For conciseness, the identifiable correlations are fixed at the same value in the simulation settings. However, none of the algorithms require this. We performed additional simulations with correlation matrices in which the identifiable

elements were not all equal. Results turn out to be very similar. One of these simulations is presented in Section A.2 of the Supplementary Materials.

For each scenario, we implemented four different methodologies: (1) the gradual rejection sampling algorithm (GRS), (2) the algorithm based on partial correlation (PC), (3) the shrinking method (SHR), and (4) the scaling method (SCA). A total of 1000 random correlation matrices were generated with each method. As a key quantity, we are interested in the computation time to draw a correlation matrix with fixed values. Furthermore, we computed the univariate (ρ_{Δ}; for $p = 1$) and multivariate (R^2_H; for $p > 1$) ICA using the generated random correlation matrices. Here, we are interested in comparing the densities of the resulting ICA quantities.

All algorithms were run on a laptop computer with an Intel(R) Core(TM) i5-6200U CPU 2.30GHz processor and 16GB of RAM.

5.2. Results

Table 1 shows the expected time to generate a random correlation matrix with fixed values using the four methods. The GRS algorithm works well for matrices of sufficiently small dimension, around eight say, and when the magnitude of the fixed correlations is sufficiently small. ~~At the same time, it is almost impossible sampling a valid correlation matrix randomly when the dimension is 12 or larger and/or when the fixed correlations are relatively large in absolute value~~. **Nevertheless, the rejection rate rapidly increases with the dimension of the matrix**. To draw a (12×12) matrix and $\rho = 0.8$, it takes more than five minutes, making it impractical. Both methods based on adjustments for non-positive-definiteness are fast, with a longer time for the SCA method. However, they do not perform well with high-dimensional matrices. The SCA method fails to keep fixed the identifiable correlations for matrices of size greater than four, even using very large weights for the fixed values. On the other hand, the shrinking parameter (α) goes to zero as the matrix size increases, leading to draws closer to 0 for the SHR method (for more details, see Section A.1 of the Supplementary Materials). Our algorithm based on partial correlations is the

fastest and generates correlations with constant symmetric densities (see Figure A.1 of the Supplementary Materials).

Table 1: Mean computation time (in seconds) to draw a correlation matrix with fixed values using different algorithms for various number of surrogates (p) and fixed ρ.

	$\rho = 0.2$				$\rho = 0.5$				$\rho = 0.8$			
p	GRS	PC	SHR	SCA	GRS	PC	SHR	SCA	GRS	PC	SHR	SCA
1	0.004	0.001	0.004	0.013	0.004	0.001	0.005	0.014	0.004	0.001	0.005	0.026
3	0.011	0.004	0.019	0.18	0.117	0.003	0.024	0.207	0.512	0.003	0.052	0.199
5	2.196	0.012	0.083	0.56	4.288	0.018	0.074	0.487	384.846	0.008	0.064	0.513
10	∞	0.023	0.283	1.592	∞	0.034	0.36	1.861	∞	0.06	0.267	1.704

The frequency densities for the ICA obtained from the different methods are displayed in Figure 1 for $p = \{1, 3, 5, 10\}$. The GRS and HP algorithm provide similar frequency densities for the ICA in all scenarios where the latter are feasible. The SHR method leads to a more peaked distribution of the ICA when p and ρ increase. The SCA algorithm provides results similar to those of the GRS and PC methods when the fixed correlations are low. However, it does not behave well for $\rho = 0.5$ or $\rho = 0.8$. Given that it fails to keep fixed the identifiable correlations, it leads to invalid values for the ICA. Note that we are using the GRS method as the reference when analyzing the ICA densities.

5.3. Identifiable bounds: Additional simulation

One of the advantages of the simulation-based sensitivity analysis introduced by [1] and [2], is that it can provide approximate identifiable bounds for the unidentifiable ICA. An additional simulation study was conducted to evaluate the performance of these identifiable bounds. The bounds are calculated based on M runs of the PC algorithm. Indeed, if the number of simulated correlation matrices is sufficiently large and the PC algorithm samples from the entire space of correlation matrices, producing the corresponding R^2_H values, then the observed $(\min R^2_H, \max R^2_H)$ should contain the true value of the ICA and, therefore, they can be considered as approximate bounds for the true ICA.

For the simulations, a setting with a bivariate $\boldsymbol{S} = (S_1, S_2)$ surrogate endpoint was considered. Three PD matrices were generated using the GRS algo-

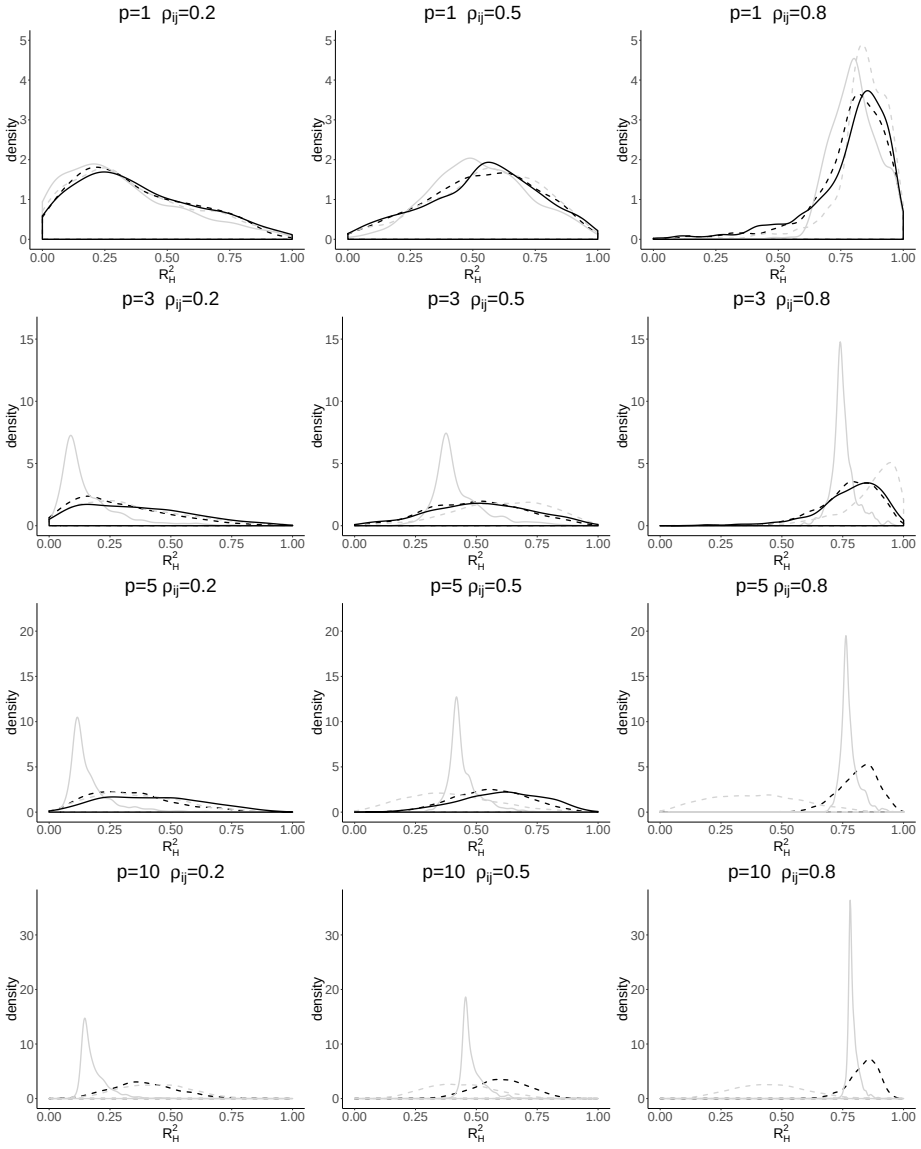

Figure 1: Density of the univariate and multivariate ICA computed using four methods to generate the unidentifiable correlations: gradual rejection sampling (solid black lines), algorithm based on partial correlations (dashed black line), linear shrinking method (solid grey line) and scaling method (dashed grey line) for different number of surrogates (p) and fixed ρ.

rithm. These matrices led to ICA values equal to 0.75, 0.85 and 0.95. Based on each of these matrices, 250 datasets of sizes $N = \{100, 200, 500\}$ were generated. When analyzing each dataset, the unidentifiable quantities were removed and the bounds for the ICA were computed by $M = \{100;\ 500;\ 1000;\ 5000;\ 10{,}000;\ 50{,}000\}$ runs of the PC algorithm. Table 2 shows the percentage of cases in which the true ICA was included in the range of the calculated R^2_H (coverage) and the average $[\min, \max]$ interval.

Table 2: Percentage of cases in which the true ICA was included in the $[\min, \max]$ (the average interval in parenthesis) range of ICA values computed using the PC algorithm.

		Number of runs (M)					
True ICA	N	100	500	1,000	5,000	10,000	50,000
0.75	100	0.54	0.68	0.75	0.83	0.84	0.88
		(0.73, 0.96)	(0.7, 0.97)	(0.69, 0.98)	(0.67, 0.98)	(0.67, 0.98)	(0.66, 0.99)
	200	0.59	0.76	0.82	0.92	0.94	0.95
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.99)
	500	0.61	0.87	0.92	0.98	0.99	1
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.99)
0.85	100	0.98	0.99	0.99	1	1	1
		(0.73, 0.96)	(0.7, 0.97)	(0.69, 0.98)	(0.67, 0.98)	(0.67, 0.98)	(0.66, 0.99)
	200	1	1	1	1	1	1
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.9)
	500	1	1	1	1	1	1
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.99)
0.95	100	0.79	0.93	0.96	0.99	0.99	1
		(0.73, 0.96)	(0.70, 0.97)	(0.69, 0.98)	(0.67, 0.98)	(0.67, 0.98)	(0.66, 0.99)
	200	0.85	0.98	1	1	1	1
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.99)
	500	0.89	1	1	1	1	1
		(0.74, 0.96)	(0.71, 0.97)	(0.70, 0.98)	(0.68, 0.98)	(0.68, 0.98)	(0.66, 0.99)

When $N = 500$ and a minimum of $M = 1000$ runs are used, valid bounds are generally obtained, i.e, the coverage probability exceeds 90% in all cases and it is often larger than 95%. However, when the ICA equals 0.75 and the sample size is relatively small and more runs may be necessary to obtained satisfactory bounds. Regarding the average $[\min, \max]$ interval, it always covers the true ICA, even

in the case of low coverage, and gets wider as the number of runs increase. Furthermore, it does not change with the true ICA. This last result is expected as the identifiable parameters are equal in the three settings. Extra simulations showed that the GRS algorithm provides reasonably similar coverage rates as the ones exhibited in Table 2, at least with $M = \{100, 200, 500\}$. The coverage evaluation for higher numbers of runs of this algorithm is computationally too time-consuming.

6. The transPAT microbiome intervention study

The TransPAT experiment [23] is an animal study conducted to evaluate the influence of an antibiotic treatment on the immune system (Immunoglobulin A level, IgA level). The dataset consists of information from 15 germ-free mice that received cecal contents of a donor mouse. The cecal contents of seven donor mice were exposed to a tylosin pulse (experimental treatment group) and eight mice were not exposed (control treatment group). 12 days after starting the experiment, the relative abundance of a total of 67 operational taxonomic units (OTUs) was measured. Regarding multiple surrogates assessment, the objective of the study was to evaluate whether the treatment effect on one or more OTUs (candidate surrogate endpoints) conveys information on the potential treatment effect on the immune response (IgA level at day 20; true endpoint). The data are available on github (see `https://github.com/blaser-lab/Paper-Ruiz-2017`) and was used by [2] to illustrate the multiple surrogates evaluation methodology.

6.1. Assessing the validity of a multivariate surrogate in the case study

[2] conducted a simulation-based sensitivity analysis where the R^2_H was computed using random correlation matrices simulated by the gradual rejection sampling algorithm (GRS). Since the number of potential surrogate endpoints is large ($p = 67$), they opted for a forward selection approach to identify the best set of surrogates. At first, a univariate analysis for each candidate surrogate is performed. Then, the one with the highest median R^2_H is kept. Afterwards,

multivariate analyses are conducted, including the first chosen surrogate combined with each of the remaining candidates. Then, the pair of surrogates with the highest median R^2_H is retained. The procedure continues until the group of selected surrogates reach a criterion, e.g., until the median $R^2_H > 0.95$. In this Section, we proceed in the same way but using the PC algorithm generating 10,000 random correlation matrices. Section A.3 of the Supplementary Materials shows how to conduct a multiple surrogacy analysis using the Surrogate R package.

Figure 2 (left hand side) shows the median R^2_H of all sets of surrogates evaluated in each step. In the univariate evaluation, not all the candidates seem to be good, several medians R^2_H are very small. Nevertheless, the median of the best one (S_1) is relatively high. As more candidates are jointly evaluated, the medians increase. In the third step, all the evaluations show a median higher than 0.85. Figure 2 (right hand side) displays the range of R^2_H values obtained from the best combination of surrogates in each step. When the best candidate surrogate is individually evaluated, the R^2_H exhibits a wide range, indicating a strong impact of the unverifiable parameters. However, the range gets narrower as more surrogates are simultaneously assessed. With five surrogates, the median R^2_H reaches 0.976 with a minimum of 0.922 and a maximum of 0.999. Until this step, the selected vector of surrogates corresponds to S_1 =OTU 44, S_2 =OTU 17, S_3 =OTU 37, S_4 =OTU 40 and S_5 =OTU 30 (S_1 belongs to the family *Ruminococcaceae*, S_2 to *Verrucomicrobiaceae*, and S_3, S_4 and S_5 are members of the *Lachnospiraceae* family).

[2] performed the forward selection approach until step three, finding fairly very similar results for the R^2_H. However, they selected a different set of surrogates after the first step. This can be explained by the fact that the bivariate analysis of most of the other candidates and S_1 provide high R^2_H values (see Figure 2, left hand side). Around 15% of the them combined with S_1 lead to a median R^2_H higher than 85%.

The median time to generate 500 correlation matrices in the evaluation of one to all surrogates is presented in Figure 3. The evaluation of a small number

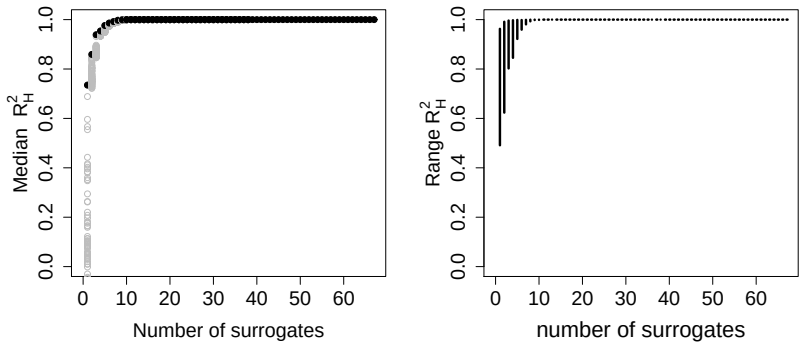

Figure 2: TransPAT data. Median R^2_H computed by the PC algorithm of all the sets of surrogates evaluated in each step of the forward selection (left hand side), where the black dot represents the highest median, and the grey dots the median of the rest of sets. $(\min R^2_H, \max R^2_H)$ interval of the best set of surrogate in each step (right hand side).

of surrogates is fast, e.g., around 3 seconds to assess five surrogates, but the computation time increases exponentially. Nevertheless, the PC algorithm can still generate large random correlation matrices in a relatively short time. The evaluation of 67 surrogates, i.e., simulating 500 (136×136) correlation matrices with 4488 fixed values, took around 56 minutes.

7. Final remarks

Various methods for generating random correlation matrices were proposed for the setting where some of the correlations are fixed, a typical situation in the causal-inference framework. We focus on multiple surrogacy assessment using the ICA in which the partially identifiable correlation matrix has a specific pattern (1). In this case, the simulation study showed that all methods exhibit reasonably similar performance with small size matrices. However, for medium to high dimensional matrices, the adaptation of the algorithm based on partial correlations (PC) outperforms the others. It is stable and computationally highly efficient.

In the multiple surrogacy evaluation, the computation of the R^2_H by the GSR

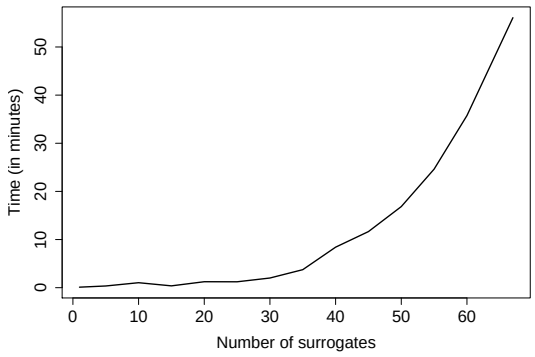

Figure 3: TransPAT data. Median computation time to calculate R^2_H based on 500 randomly generated correlation matrices using the PC algorithm.

method is computationally intensive even with few surrogates, and unfeasible with more than five, limiting the joint evaluation to a relatively small number of candidates. On the other hand, the PC algorithm allows assessing more surrogates in a short time. Although the evaluation of a large number of surrogates combined is not commonly encountered in practical cases, the PC algorithm is not only limited to this framework. It can be used with high-dimensional matrices. In the case study, we were able to simulate random correlation matrices of size (136×136) (evaluating all candidates combined) in a relatively fast manner. Furthermore, it allows incorporating additional restrictions about the unidentifiable correlations easily, e.g., biologically implausible values. However, this may affect the marginal distribution of the single correlations considerably.

The PC algorithm can be implemented to a different arrangement of known entries of the correlation matrix by reindexing its columns and rows. However, it does not mean that this method is the best, or even feasible, for all patterns. Depending on the positioning of the fixed values in the matrix, a different choice of partial correlations may lead to computationally better results. For more details on this, we refer to [5, 6] and [4].

Acknowledgements

Financial support from the IAP research network $\#P7/06$ of the Belgian Government (Belgian Science Policy) is gratefully acknowledged. Alvaro J. Flórez acknowledges funding from the European Seventh Framework programme $FP7$ $2007-2013$ under grant agreement Nr. 602552.

Supplementary Materials

Additional simulations, and an illustration of the R package Surrogate can be found online.

References

[1] A. Alonso, W. Van der Elst, G. Molenberghs, M. Buyse, T. Burzykowski, On the relationship between the causal-inference and meta-analytic paradigms for the validation of surrogate endpoints, Biometrics 71 (1) (2015) 15–24.

[2] W. Van der Elst, A. Alonso, H. Geys, P. Meyvisch, L. Bijnens, R. Sengupta, G. Molenberghs, Univariate versus multivariate surrogate endpoints in the single-trial setting, Statistics in Biopharmaceutical Research 0 (0) (2019) 1–17.

[3] H. Joe, Generating random correlation matrices based on partial correlations, Journal of Multivariate Analysis 97 (10) (2006) 2177–189.

[4] D. Lewandowski, D. Kurowicka, H. Joe, Generating random correlation matrices based on vines and extended onion method, Journal of Multivariate Analysis 100 (9) (2009) 1989–2001.

[5] D. Kurowicka, R. M. Cooke, A parameterization of positive definite matrices in terms of partial correlation vines, Linear Algebra and its Applications 372 (2003) 225–251.

[6] D. Kurowicka, R. M. Cooke, Completion problem with partial correlation vines, Linear Algebra and its Applications 418 (1) (2006) 188–200.

[7] M. Pourahmadi, X. Wang, Distribution of random correlation matrices: Hyperspherical parameterization of the cholesky factor, Statistics & Probability Letters 106 (2015) 5–12.

[8] P. I. Davies, N. J. Higham, Numerically stable generation of correlation matrices and their factors, BIT Numerical Mathematics 40 (4) (2000) 640–651.

[9] M. Mittelbach, B. Matthiesen, E. A. Jorswieck, Sampling uniformly from the set of positive definite matrices with trace constraint, IEEE Transactions on Signal Processing 60 (5) (2012) 2167–2179.

[10] V. Madar, Direct formulation to cholesky decomposition of a general non-singular correlation matrix, Statistics & Probability Letters 103 (2015) 142–147.

[11] P. J. Rousseeuw, G. Molenberghs, Transformation of non positive semidefinite correlation matrices, Communications in Statistics, Theory and Methods 22 (4) (1993) 965–984.

[12] N. Higham, N. Strabić, V. Šego, Restoring definiteness via shrinking, with an application to correlation matrices with a fixed block, SIAM Review 58 (2) (2016) 245–263.

[13] T. Burzykowski, G. Molenberghs, M. Buyse, The Evaluation of Surrogate Endpoints, Springer-Verlag GMBH, 2005.

[14] G. Molenberghs, T. Burzykowski, A. Alonso, P. Assam, A. Tilahun, M. Buyse, A unified framework for the evaluation of surrogate endpoints in mental-health clinical trials, Statistical Methods in Medical Research 19 (3) (2010) 205–236.

[15] A. Alonso, T. Bigirumurame, T. Burzykowski, M. Buyse, G. Molenberghs, L. Muchene, N. Perualila, Z. Shkedy, W. Van der Elst, Applied Surrogate Endpoint Evaluation with SAS and R, Chapman & Hall/CRC, Boca Ratón, 2016.

[16] D. B. Rubin, Statistics and causal inference: Comment: which ifs have causal answers, Journal of the American Statistical Association 81 (396) (1986) 961–962.

[17] P. W. Holland, Statistics and causal inference, Journal of the American Statistical Association 81 (396) (1986) 945–960.

[18] P. R. Rosenbaum, D. B. Rubin, The central role of the propensity score in observational studies for causal effects, Biometrika 70 (1) (1983) 41–55.

[19] W. Van der Elst, G. Molenberghs, A. Alonso, Exploring the relationship between the causal-inference and meta-analytic paradigms for the evaluation of surrogate endpoints, Statistics in Medicine 35 (8) (2015) 1281–1298.

[20] E. Linfoot, An informational measure of correlation, Information and Control 1 (1) (1957) 85–89.

[21] H. Joe, Relative entropy measures of multivariate dependence, Journal of the American Statistical Association 84 (405) (1989) 157–164.

[22] W. Böhm, K. Hornik, Generating random correlation matrices by the simple rejection method: Why it does not work, Statistics & Probability Letters 87 (2014) 27–30.

[23] V. E. Ruiz, T. Battaglia, Z. D. Kurtz, L. Bijnens, A. Ou, I. Engstrand, X. Zheng, T. Iizumi, B. J. Mullins, C. L. Mller, K. Cadwell, R. Bonneau, G. I. Perez-Perez, M. J. Blaser, A single early-in-life macrolide course has lasting effects on murine microbial network topology and immunity, Nature Communications 8 (1).