NORTA method using numerical integration and Newton's algorithm

May 2020

1 Problem Statement

Let $X_1, ..., X_n$ be a collection of n random variables with Pearson correlation matrix Σ_X, finite variance $var(X_i)$ and finite expected value $E[X_i]$, each following a distribution with cdf $F_i(x)$ $\forall i \in \{1, 2, ..., n\}$. We are interested in drawing a random sample from $(X_1, ..., X_n)$ having the desired correlation matrix Σ_X. Note that the specified marginals and correlation matrix are not enough to determine the joint probability distribution of $(X_1, ..., X_n)$, so the probability integral transformation method cannot be directly applied.

2 Important results and formulas

Consider the transformation $X_i = F_i^{-1}(\Phi(Z_i)$ where Z_i is the ith element of a standard random multivariate normal vector $(Z_1, ..., Z_n)$ with correlation matrix Σ_Z

Recall that Pearson correlation between two random variables X_i and X_j with finite variance and expected value is defined as:

$$\frac{E[X_i X_j] - E[X_i]E[X_j]}{\sqrt{var(X_i)var(X_j)}} \tag{1}$$

As variance and expected value are fixed (since we have control on the parameters of every marginal), the expression above depends entirely on the expected value of the random variable $X_i X_j = F_i^{-1}(\Phi(Z_i))F_j^{-1}(\Phi(Z_j))$. The formula for the expected value of such variable is

$$E[X_i X_j] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F_i^{-1}(\Phi(z_i)) F_j^{-1}(\Phi(z_j)) \varphi_{\Sigma_Z(i,j)}(z_i, z_j) \, dz_i dz_j \tag{2}$$

where $\varphi_{\Sigma_Z(i,j)}$ is the standard bivariate normal pdf with correlation $\Sigma_Z(i,j)$

The following propositions are proved in Cario & Nelson, where $c_{ij}(\Sigma_Z(i,j))$ is the expression in (1) as a function of the i,jth entry of the correlation matrix Σ_Z:

Proposition 1 *The function $c_{ij}(\Sigma_Z(i,j))$ is a non decreasing function of the i,j-th entry of Σ_Z.*

Proposition 2 *Let $p_{\max}(i,j)$ and $p_{\min}(i,j)$ be the maximum and minimum feasible Pearson correlation between random variables X_i and X_j, for $i \neq j$, then*

$$\lim_{\Sigma_Z(i,j)\to+1} c_{ij}(\Sigma_Z(i,j)) = p_{\max}(i,j)$$

and

$$\lim_{\Sigma_Z(i,j)\to-1} c_{ij}(\Sigma_Z(i,j)) = p_{\min}(i,j)$$

3 NORmal To Anything

The following procedure was introduced by Cario & Nelson (1997):

To construct a random sample from $(X_1, ..., X_n)$, firstly we generate an n-dimensional standard normal random vector Z with Pearson correlation matrix Σ_Z and then transform it into a vector U of standard uniform random variables by using the integral transformation method; such vector has a Gaussian copula distribution. After this, by the inverse transformation technique[1], we transform the uniform random vector U back into a random vector X with the target marginals. The first step in the NORTA algorithm requires us to find the correlation matrix Σ_Z of the normal random vector such that it guarantees the vectors generated have a specific Pearson correlation matrix. To reach this goal we solve $\frac{n(n-1)}{2}$ one-dimensional simultaneous equations, which usually is difficult to do analytically (though in the examples given by Cario and Nelson they use a closed form of the functions c_ij). Moreover, there are two complications in NORTA algorithm. The first is that at least one of the desired correlations may not be feasible; that is, there may not exist a correlation $\Sigma_Z(i,j)$ such that the corresponding transformed random variables X_i and X_j have the desired correlation value of $\Sigma_X(i,j)$. The second restriction is that even if all entries of Σ_X are feasible, the correlation matrix Σ_Z for the multivariate normal random vector Z may not be positive definite. This becomes an issue where the dimensionality of the random vector increases (Ghosh & Henderson, 2003).

4 The algorithm

- Input: The cdf of each marginal $F_1, ..., F_n$, a correlation matrix Σ_X and N, the number of sample points to simulate.
- Output: An $N \times n$ matrix with each row being an observation of the random vector $(X_1, ..., X_n)$ having a correlation matrix of Σ_X
- Step 1: For each i and $j \in \{1, ..., n\}$ such that $i < j$:

[1] `https://stats.stackexchange.com/questions/184325/how-does-the-inverse-transform-method-work`

Compute a numerical version of the function $c_{ij}(\Sigma_Z(i,j))$ by using numerical integration methods.

Evaluate $c_{ij}(\Sigma_Z(i,j))$ at the limits ± 1 and break the process if at least one entry of Σ_X is not feasible

Use Newthon's method[2] to find the root of the equation $c_{ij}(\Sigma_Z(i,j)) - \Sigma_X(i,j)$

- Step 2: Construct a matrix Σ_Z by using the roots found in previous step and check whether is positive definite. If it is not, solve a SDP to find a positive definite matrix that is near to Σ_Z
- Step 3: Generate a sample of size N from a standard multivariate random normal vector $Z = (Z_1, ..., Z_n)$ with correlation matrix Σ_Z
- Step 4: Transform the sample in step 3 into a sample of standard uniform marginals by using the transformation $U_i = \Phi(Z_i)$
- Step 5: Transform the sample in Step 4 into a sample of NORTA vectors $X = (X_1, ..., X_n)$ by using the transformation $X_i = F_i^{-1}(U_i)$. This final sample is the desired output

Note that in order to overcome the first complication mentioned in the previous section, we use proposition 2 to evaluate the feasibility of each pairwise correlation and check if Σ_X is feasible. As for the second complication, the nearest positive definite matrix algorithm is used to find a correlation matrix that is positive definite and nearest to the calculated non-positive definite matrix Σ_Z.

5 Comments

It seems that the algorithm works very well when correlating continuous marginals. It still works quite well when dealing with discrete variables, though the error increases.

A natural question arises at the moment of using this algorithm. If there exists a random vector $(X_1, ..., X_n)$ with correlation matrix Σ_X, does this method always lead to such vector? Ghosh and Henderson (2003) showed that there are some cases when NORTA fails, and then studied the problem as n increases. As we pointed out, some feasible correlation matrices Σ_X could lead to a non-positive definite intermediate matrix Σ_Z and hence NORTA is unable to output a random vector with the desired properties. We fixed this by using the nearest PD matrix, which is also what Ghosh and Henderson studied and proposed.

[2] Proposition 1 and continuity ensure that there is at least one solution (intermediate value theorem)

6 Bibliography

Ghosh, S., Henderson, S. G. (2003). Behavior of the NORTA method for correlated random vector generation as the dimension increases. ACM Transactions on Modeling and Computer Simulation, 13(3), 276–294. https://doi.org/10.1145/937332.937336

Matrix Marne, C., Cario Delphi, Electric, P., Systems Warren (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation.

Niavarani, Smith, A. (2013). Modeling and generating multi-variate-attribute random vectors using a new simulation method combined with NORTA algorithm.