Algorithm feasibility analysis

Coordinator: Janez Komelj

1 Key information

Assumptions: Let $\mathbf{X} = (X_1, \ldots, X_n)^t$ be a random column vector. Its components are continuous random variables with known cumulative distribution functions $F_1, \ldots, F_n$, finite mathematical expectations $\mu_1, \ldots, \mu_n$ and finite variances $\sigma_1^2, \ldots, \sigma_n^2$. Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be the diagonal of the diagonal matrix $\mathbf{D}$. Assuming that $X_1, \ldots, X_n$ are dependent and their covariance matrix Σ is nonsingular, the correlation matrix $\mathbf{P} = \mathbf{D}^{-1}\Sigma\mathbf{D}^{-1}$ is positive definite.

Problem: Given $F_1, \ldots, F_n$ and a positive definite correlation matrix $\mathbf{P}$, generate such a $n \times m$ dimensional matrix $\mathbf{A}$ with a corresponding correlation matrix $\hat{\mathbf{P}}$ that the columns of $\mathbf{A}$ are random samples of $\mathbf{X}$, and $\hat{\mathbf{P}}$ is equal or close to $\mathbf{P}$.

Goal: After solving the problem, we can calculate a marginal sum of $\mathbf{A}$ by rows and it is a sample of $S = \sum_{i=1}^n X_i$. Its empirical cdf and quantile function can be used as an approximation for an unknown F_S and F_S^{-1}, respectively.

Open challenge: Given are $F_1, \ldots, F_n$, $\mathbf{P}$ and an algorithm in the sequel which iteratively calculates matrices $\mathbf{A}_k$ and corresponding $\hat{\mathbf{P}}_k$. Assuming an unlimited number of iterations, select any matrix norm and find a sound theoretical foundation to answer the questions:

- Under which conditions the sequence $\hat{\mathbf{P}}_0, \hat{\mathbf{P}}_1, \hat{\mathbf{P}}_2, \ldots$ converges to $\mathbf{P}$;
- Under which conditions the algorithm does not stop, but the matrices $\hat{\mathbf{P}}_k$ for large enough k and reasonable $\varepsilon > 0$ fall into the ε-neighborhood of $\mathbf{P}$ and stay there;

What is expected: Formulation of a theorem with proof.

What is not expected: Proposals for improvements of algorithm regarding space and/or time complexity are not expected. However, if a change in the algorithm improves its feasibility and/or the speed of convergence, it is desired. Such a case could appear, if $\mathbf{P}$ is barely positive definite. In any case, a theoretical underpinning of the improvement is preferred to a statistical one, and the expectations are the same, but for the modified algorithm.

2 Insurance and mathematical background of the problem

$X_1, \ldots, X_n$ can be seen as dependent risks in insurance for which the correlation matrix $\mathbf{P}$ is usually not known, but is prescribed in legislation. We are interested in the cdf F_S which enables better risk management and calculations of capital requirements. More details are in the appendix.

For given $F_1, \ldots, F_n$ and $\mathbf{P}$, a corresponding n-dimensional cdf $F_\mathbf{X}$ does not always exist (see McNeil, Frey, & Embrechts, 2005, p. 205, Example 5.26). According to Wang, theoretically there might exist infinitely many of them, but finding even one can be practically difficult.

In general, the continuous marginal cdfs and copulae uniquely determine $F_\mathbf{X}$ (see McNeil et al., 2005, p. 186, Theorem 5.3 (Sklar 1959)). Let us focus on random vectors whose components are connected by a normal copula which is uniquely determined by the correlation matrix $\mathbf{P}$. Consequently, in this special case $F_\mathbf{X}$ is uniquely defined by $F_1, \ldots, F_n$ and $\mathbf{P}$, but note that $\mathbf{P}$ has the role of a parameter. If the marginal cdfs $F_1, \ldots, F_n$ are also normal, then $\mathbf{X}$ is n-dimensionally normally distributed and the corresponding correlation matrix $\hat{\mathbf{P}}$ equals to $\mathbf{P}$. Generally, however, $\hat{\mathbf{P}} \neq \mathbf{P}$, because linear (Pearson's) correlation coefficients do not depend only on the copula, as is the case with Spearman's correlation coefficients of rank and Kendall's τ, but also on marginal distributions.

Let $\mathscr{P}$ be a set of all $n \times n$ dimensional correlation matrices. For given $F_1, \ldots, F_n$, let a function $f\colon \mathscr{P} \to \mathscr{P}$ be defined by $f\colon \mathbf{P} \mapsto \hat{\mathbf{P}}$. From the above-mentioned Example 5.26 follows that the image of f can be a proper subset of $\mathscr{P}$. In other words, the problem of how to find such $\tilde{\mathbf{P}}$ that $f\colon \tilde{\mathbf{P}} \mapsto \mathbf{P}$ may have no solution.

3 Solving the problem by simulation

Let us specify Algorithm 1 which is an adaptation of the algorithm from (Wang, 1998, p. 891) or a combination of algorithms 3.2 and 5.9 from (McNeil et al., 2005, p. 66 and 193). For the variant of the problem with the Student t copula instead of the normal copula, see algorithms 3.10 and 5.10 from (McNeil et al., 2005, p. 76 and 193).

Algorithm 1: *Generating m dependent random samples of the random vector $\mathbf{X} = (X_1, \ldots, X_n)^t$ with marginal cdfs $F_1, \ldots, F_n$. The components are connected by a normal copula defined by a $n \times n$ dimensional positive definite correlation matrix $\mathbf{P}$. The function returns a $n \times m$ dimensional matrix whose columns are samples of $\mathbf{X}$ while its marginal sum by rows is a sample of $S = \sum_{i=1}^{n} X_i$.* *Adapted from (Wang, 1998, p. 891)*

1. **function** GENDEP(m,$(F_1, \ldots, F_n)$,$\mathbf{P}$)
2. Perform a Cholesky decomposition of $\mathbf{P} = \mathbf{L}\mathbf{L}^t$, where $\mathbf{L}$ is a lower triangular matrix.
3. For $Z \sim \mathrm{N}[0,1]$ generate mn independent random values and compose a $n \times m$ dimensional matrix $\mathbf{Z}$.
4. Set $\mathbf{Y} = \mathbf{L}\mathbf{Z}$.
5. Set a matrix $\mathbf{U}$ with elements $u_{ij} = \Phi(y_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.
6. Set a matrix $\mathbf{A}$ with elements $a_{ij} = F_i^{-1}(u_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, m$.
7. **return** $\mathbf{A}$.
8. **end function**

Suppose that for a given $\mathbf{P}$ we calculated the matrix $\mathbf{A}$ with Algorithm 1, and the corresponding correlation matrix $\hat{\mathbf{P}}$ with the function COR($\mathbf{A}$). If the error $||\hat{\mathbf{P}} - \mathbf{P}||$, measured by the selected matrix norm, is acceptable, the marginal sum of the matrix $\mathbf{A}$ by rows is considered to be a m-dimensional sample of S. Using its empirical cdf, quantile calculation is simple.

An alternative, and often more accurate, option for the calculation of $\mathbf{A}$ is the Iman-Conover method (see Mildenhall, 2006, p. 145). The algorithm is a bit more complex, and its theoretical analysis even more so. Even Iman and Conover (1982) did not theoretically underpin it in all aspects. A sound theoretical foundation is based on Vitale's theorem (see Mildenhall, 2006, p. 185–188, and Vitale, 1990, p. 465, Theorem 3). In practice, it is often enough to check that the result is correct, but a problem emerges when it is not, and there is no alternative.

4 Iterative problem solving with a parameter adjustment

Assume that Algorithm 1 responds to a small change in parameter $\Delta\mathbf{P}$ by a small change in the result $\Delta\hat{\mathbf{P}} \approx \Delta\mathbf{P}$. When we get $\hat{\mathbf{P}}_0$ and a too large error $\hat{\mathbf{P}}_0 - \mathbf{P}$ with the parameter $\mathbf{P}_0 = \mathbf{P}$, in the next step we calculate with an adjusted parameter $\mathbf{P}_1 = \mathbf{P}_0 - (\hat{\mathbf{P}}_0 - \mathbf{P}) = \mathbf{P} - (\hat{\mathbf{P}}_0 - \mathbf{P}_0)$, getting $\hat{\mathbf{P}}_1$ and $\hat{\mathbf{P}}_1 - \mathbf{P}$. If necessary, we continue with $\mathbf{P}_2 = \mathbf{P}_1 - (\hat{\mathbf{P}}_1 - \mathbf{P}) = \mathbf{P} - (\hat{\mathbf{P}}_1 - \mathbf{P}_1)$ etc. The described procedure is built into Algorithm 2.

Algorithm 2: *Generating m dependent random samples of the random vector* $\mathbf{X} = (X_1, \ldots, X_n)^t$ *with marginal cdfs* $F_1, \ldots, F_n$*. The components are connected by a normal copula defined by a* $n \times n$ *dimensional positive definite correlation matrix* $\mathbf{P}$*. The function returns a* $n \times m$ *dimensional matrix whose columns are samples of* $\mathbf{X}$ *while its marginal sum by rows is a sample of* $S = \sum_{i=1}^{n} X_i$*.* *Janez Komelj*

1. **function** ITERGENDEP$(m,(F_1,\ldots,F_n),\mathbf{P},\varepsilon,k_{max})$
2. $\quad k := 0$
3. $\quad \mathbf{P}_0 := \mathbf{P}$
4. $\quad \mathbf{A}_0 :=$ GENDEP$(m,(F_1,\ldots,F_n),\mathbf{P}_0)$ ▷ Assume that the random number generator seed is initialized.
5. $\quad \hat{\mathbf{P}}_\mathbf{0} :=$ COR$(\mathbf{A}_0)$
6. $\quad$ **while** $||\hat{\mathbf{P}}_k - \mathbf{P}|| > \varepsilon$ **and** $k < k_{max}$ **do**
7. $\qquad k := k+1$
8. $\qquad \mathbf{P}_k := \mathbf{P} - (\hat{\mathbf{P}}_{k-1} - \mathbf{P}_{k-1})$
9. $\qquad \mathbf{A}_k :=$ GENDEP$(m,(F_1,\ldots,F_n),\mathbf{P}_k)$ ▷ Assume that the random number generator seed is initialized.
10. $\qquad \hat{\mathbf{P}}_\mathbf{k} :=$ COR$(\mathbf{A}_k)$
11. $\quad$ **end while**
12. $\quad$ **return** $\mathbf{A}_k$
13. **end function**

In Algorithm 2, the function GENDEP must be deterministic in terms of returning the same results for the same arguments regardless of where and when it is called. Therefore, suppose that we initially save the current seed of the generator of the (pseudo) random numbers, initialize it, and, finally, restore it.

Algorithm 2 has been tested, but not in insurance practice. However, the author has very good experience with an analogous algorithm, written in the programming language R, where the Iman-Conover method is used instead of Algorithm 1. For example, for $n = 12$, $m = 200{,}000$, the Frobenius matrix norm, and $\varepsilon = 10^{-6}$, neither convergence nor execution time were problematic. However, the accuracy is exaggerated, because all correlation matrices were very crude estimates – from the Solvency 2 documents where all off-diagonal elements are from the set $\{-\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\}$.

Less accurate results were usually related to the nature of the problem where the prescribed linear correlations are simply not achievable for the given marginal cdfs (see Embrechts, McNeil, & Straumann, 1999). Serious problems appeared only in cases where matrix $\mathbf{P}$ was barely positive definite, but $\mathbf{P}_1$ or any later one was not, which happened when modeling some financial risks. It was, however, sometimes possible to bypass the problem by modifying step 8.

Expecting that practical experience would not be significantly worse with the simpler Algorithm 2, the following task is set.

Assuming an unlimited number of iterations, theoretically analyze Algorithm 2, determine and prove:

- Under which conditions do we obtain a sequence of matrices $\mathbf{A}_k$ for which the corresponding correlation matrices $\hat{\mathbf{P}}_\mathbf{k} = \text{COR}(\mathbf{A}_k)$ converge to a prescribed correlation matrix $\mathbf{P}$.
- Under which conditions the algorithm does not stop, but the matrices $\hat{\mathbf{P}}_k$ for large enough k and $\varepsilon \in (0,0.05]$ fall into the ε-neighborhood of $\mathbf{P}$ and stay there.

You can expand the task and find out more, correct it if it is poorly defined, or adapt it. For example, the interval for ε is calibrated for the maximum matrix norm. It is ideal only for a stop condition, but not otherwise, because $||\mathbf{AB}|| \leq ||\mathbf{A}||\,||\mathbf{B}||$ does not hold.

If the task is too easy, replace the normal copula with the Student t copula and/or replace Algorithm 1 with the Iman-Conover algorithm. Both modifications require only minor changes in Algorithm 2.

References

1. Embrechts, P., McNeil, A., & Straumann, D. (1999). Correlation: Pitfalls and Alternatives. *Risk*, *12*(5), 69–71.
2. Iman, R. W., & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics – Simulation and Computation*, *11*(3), 311–334.
3. McNeil, A. J., Frey, R., & Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton: Princeton University Press.
4. Mildenhall, S. J. (2006). The Report of the Research Working Party on Correlations and Dependencies Among All Risk Sources – Part 1: Correlation and Aggregate Loss Distributions with an Emphasis on the Iman-Conover Method. In *Casualty Actuarial Society Winter Forum* (pp. 103–203). Arlington: Casualty Actuarial Society.
5. Vitale, R. A. (1990). On Stochastic Dependence and a Class of Degenerate Distributions. In H. W. Block, A. R. Sampson, & T. H. Savits (Eds.), *Topics in statistical dependence* (Vol. 16, pp. 459–469). Hayward: Institute of Mathematical Statistics.
6. Wang, S. S. (1998). Aggregation of Correlated Risk Portfolios: Models and Algorithms. In *Proceedings of the Casualty Actuarial Society* (Vol. 85, pp. 849–939). Arlington: Casualty Actuarial Society.

Appendix – insurance background of the problem

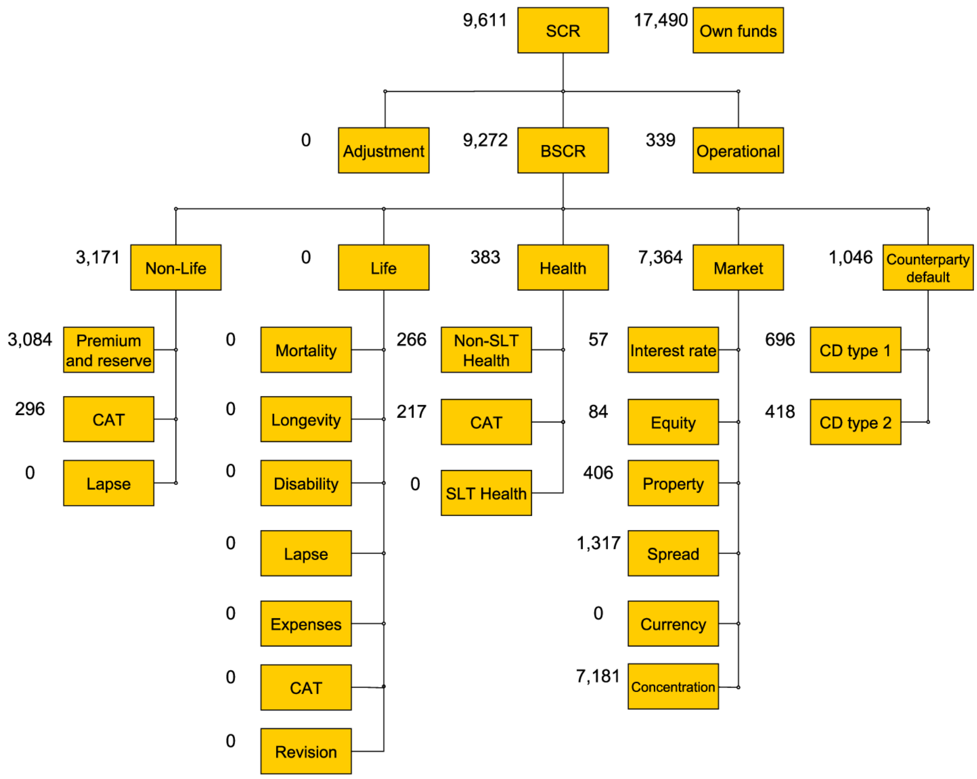
The new European insurance legislation Solvency 2 entered into force in 2016. Our task is related to the following important points introduced by it:

- New capital requirements for EU (re)insurers;
- Risk based capital system;
- Standard formula versus internal model;
- Enterprise risk management;

Many hierarchically structured risks need to be considered, and for each of them such (partial) solvency capital requirements (SCR) must be calculated, that they are neutralized, taking into account 99.5% probability and the time horizon of one year. At the highest level, we get a key requirement of Solvency 2: The probability that the insurer remains solvent for one year must be at least 99.5%.

The standard formula is provided for the calculation of the SCR, and it also prescribes what are admissible own funds which must be at least as large as the SCR. An example of the SCR structure is shown in Figure 1.

Figure 1: SCR structure for the hypothetical non-life insurance company (€ 1,000)



Note that at the highest level the SCR equals to $BSCR + SCR_{operational} - Adjustment$ (details are irrelevant here), but at the lower levels the SCRs are not the sum of subordinate items. They are calculated bottom up by a square root formula, taking into account the prescribed risks correlations. The basic SCR is calulated by

$$BSCR = \sqrt{\sum_{i,j} \rho_{ij} \times SCR_i \times SCR_j},$$

where running indexes i and j must take all possible values (non-life, life, health, market and counterparty default). The correlation coefficients ρ_{ij} are elements of the prescribed correlation matrix in Table 1.

Table 1: Basic Solvency Capital Requirement – correlation matrix

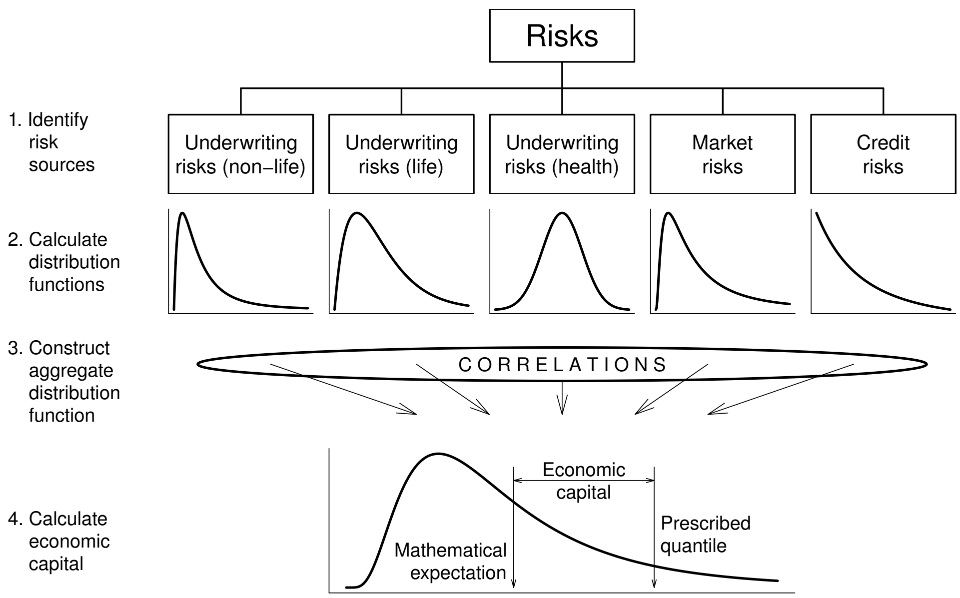
	Non-life	Life	Health	Market	Default
Non-life	1	0	0	0.25	0.50
Life	0	1	0.25	0.25	0.25
Health	0	0.25	1	0.25	0.25
Market	0.25	0.25	0.25	1	0.25
Default	0.50	0.25	0.25	0.25	1

Source: Annex IV of Directive 138/2009/EC

If the insurance company considers the standard formula unsuitable for it, it can build its own internal or partial internal model. However, such a model must be approved by the insurance supervisor, implying that there exists a sound theoretical background, a long history of data, a good quality of data, and a lot of other conditions which must also be fulfilled.

In principle, the *BSCR* calculation could be done as sketched in Figure 2. Since the SCR calculation process goes bottom up, the distribution functions in the second step, which are an input for the fourth step, can be calculated analogously.

Figure 2: Solvency Capital Requirement calculation under Solvency 2



Source: Adapted from Stuart Wason: Insurer Solvency Assessment: Towards a Global Framework, 2004

If the cdf F_S of the risk S is known, the SCR calculation reduces to $SCR(S) = F_S^{-1}(0.995) - \mathbb{E}[S]$. It might appear easy in theory, but it is not in practice.