Calibrating the dependence structure of the CreditRisk$^+$ model at different time scales

Jacopo Giacomelli*

SACE S.p.A,
Piazza Poli 37/42, 00187, Rome (Italy)
j.giacomelli@sace.it

Luca Passalacqua†

Sapienza, University of Rome, Department of Statistics,
Viale Regina Elena 295, 00161 Rome (Italy)
luca.passalacqua@uniroma1.it

July 6, 2018

Abstract

The CreditRisk$^+$ model is one of the industry standards for the valuation of credit loans portfolios or credit insurance policies. The calibration of CreditRisk$^+$ model requires, *inter alia*, the specification of the parameters describing the structure of dependence of default events, loosely speaking "default correlations", that – using copula functions – can be shown to correspond to a multivariate Clayton copula. This work addresses the calibration of the structure of dependence. In particular, we study the dependence of the calibration procedure on the tenor of the time series, that might have a different frequency with respect to the time horizon onto which the model is used for forecasting, as it is often the case in real life applications. The role of the statistical error as a function of the time series tenor is also discussed.

*The views and opinions expressed in this article are those of the author and do not necessarily reflect the official policy or position of SACE S.p.A.

†corresponding author.

1 Introduction

While the development of modern portfolio models started within the framework of the Basel Accords, it is with the great credit crisis of 2008, triggered by the subprime mortgage crisis, that increasing attention started to be paid to the role of the structure of dependence, loosely speaking the "correlations", between defaut events. In fact, it is well established that tails of the distribution of the value of asset/liabilities portfolios are in fact dominated by the structure of dependence, rather than by the values of the other basic components of credit risk, namely the Probability of Default (PD), the Exposure At Default (EAD) and the Loss Given Default (LGD). In this regard, the calibration issues raised by the choice of a particular structure of dependence can be as important as the choice of the model itself. Calibration of the structure of dependence of a portfolio model is – in general – a very demanding task, since the number of parameters should be large, in order to cope with the complexity of real data, and, on the other hand, data are usually not numerous enough to fill the sample space in a way sufficient for a precise estimation of the parameters.

In this work we address a typical real life problem, namely how to chose the frequency of the historical time series used to calibrate a classic credit portofilo model, CreditRisk$^+$, in order to provide the most accurate estimation of the structure of dependence parameters, or, in other words, how the calibration precision *scales* with the time series frequency.

CreditRisk$^+$, born 1997, is at present still one of the financial and actuarial industry standards for the assessment of credit risk in portfolios of financial loans or credit/suretyship policies. After introducing the model, the following sections will present a modified version of it, that will be later useful in deriving the main results of the work, *i.e.* how to calibrate the structure of dependence of the model with historical time series of arbitrary frequency and how the statistical precision depends on the frequency. Finally, numerical examples will be used to illustrate the analytic results.

2 The CreditRisk$^+$ model

The CreditRisk$^+$ model is a portfolio model developed by Credit Suisse First Boston (CSFB) by Tom Wilde and coworkers, firstly documented in [1] and later widely discussed in [2]. It is a model *actuarially inspired* in the sense that losses are due only to default events and not to other sources of financial risk, like *e.g.* variation of the credit standing (the so-called "credit migration" effect). In fact, CreditRisk$^+$ can be classified as a *frequency-severity model* with the peculiarity that the frequency of default events is described by a doubly stochastic process while – at least in the original formulation of the model –loss severity is deterministic. The second hypothesis can be easily relaxed at the cost of some additional computational burden. However, this issue can be negleted for what follows.

Another particular feature of the model is that default events are correlated, which makes the model suited both for financial applications and for actuarial applications to lines of business such as credit or suretyship insurance. The structure of dependence of default events is described using a factor model framework, where factors are unobservable ("latent") stochastic "market" variables, whose precise financial/actuarial identification is irrelevant, since the model integrates on all possibile realizations ("market scenarios"). Therefore, CreditRisk$^+$ can be further classified inside the familty of *factor models*, and in particular inside the sub-family of *conditionally independent* factor models, since, conditionally on the values assumed by the factors, defaults are supposed (by the model) to be independent.

The structure of the model can be summarised as follows. Let N be the number of different risks in a given portfolio and $\mathbb{1}_i$ the default indicator function of the i-th risk ($i = 1, \ldots, N$) over the time horizon $(t, T]$. The indicator function $\mathbb{1}_i$ is a Bernoulli random variable that takes the value 1 in case of default with probability q_i and the value 0 with probability $1 - q_i$. Thus:

$$\mathbf{E}\left[\mathbb{1}_i\right] = q_i, \quad \mathbf{cov}\left[\mathbb{1}_i\right] = q_i(1 - q_i), \quad i = 1, \ldots, N. \tag{1}$$

The "portfolio loss" L over the reference time horizon (t, T) is then given by:

$$L = \sum_{i=1}^{N} \mathbb{1}_i \, E_i, \quad E_i = (EAD)_i \, (LGD)_i, \tag{2}$$

where $(EAD)_i$ and $(LGD)_i$ are respectively the *Exposure At Default* and the *Loss Given Default* of the i-th risk, while E_i will be briefly indicated as the *Exposure* of the i-th risk, which is supposed to be deterministic.

In order to ease the semi-analytic computation of the distribution of L, the model introduces a new set of variables Y_i, each replacing the corresponding indicator function $\mathbb{1}_i$ ($i = 1, \ldots, N$). The new variables Y_i are supposed to be Poisson-distributed, conditionally on the value assumed by the market latent variables.

Let K be the number of latent variables and $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ the K-dimensional latent variable describing the "market". The latent variables are assumed to be

independent from each other and gamma-distributed, so that their joint distribution is:

$$f(\boldsymbol{x}) = \prod_{k=1}^{K} \frac{x_k^{\alpha_k - 1}}{\beta_k^{\alpha_k}\Gamma(\alpha_k)} e^{-x_k/\beta_k}, \quad x_k \geq 0, \quad \alpha_k, \beta_k > 0, \tag{3}$$

with expected value and covariance matrix given by:

$$\mathbf{E}[\boldsymbol{\Gamma}] = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix}, \quad \mathbf{cov}[\boldsymbol{\Gamma}] = \Sigma = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_{\mathrm{K}}^2, \end{pmatrix}, \tag{4}$$

where $\mu_k = \alpha_k \beta_k$ and $\sigma_k^2 = \alpha_k \beta_k^2 = \mu_k \beta_k$.
Without loss of generality, in the following it will be assumed that

$$\mu_1 = \dots = \mu_{\mathrm{K}} = 1, \tag{5}$$

so that:

$$\alpha_k = \beta_k^{-1}, \quad \beta_k = \sigma_k^2, \quad k = 1, \dots, K. \tag{6}$$

Moreover, given that Σ is diagonal, $\boldsymbol{\sigma}_\Gamma^2 = \mathrm{diag}(\Sigma)$ will be used hereafter.

Remark. In the original formulation of the model there is no information on the time evolution of the latent variables. It is interesting to note that variables evolving according to a mean-reverting square-root process, that is suited to model economic cycles, are asymptotically gamma-distributed. ■

Conditionally on the value taken by the latent variables, the parameter of the Poisson distribution of the Y_i variable is assumed to be:

$$p_i(\boldsymbol{\Gamma}) = q_i \cdot \left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik} \Gamma_k \right), \tag{7}$$

where the *factor loadings* ω_{ik} are supposed to be all non-negative and to sum up to unity:

$$\begin{aligned} &\omega_{ik} \geq 0 \qquad && i = 1, \dots, N, \quad k = 0, \dots, K, \\ &\sum_{k=0}^{K} \omega_{ik} = 1, && i = 1, \dots, N, \end{aligned} \tag{8}$$

so that q_i is the unconditional expected default frequency:

$$q_i = \mathbf{E}[p_i(\boldsymbol{\Gamma})] = \int_0^\infty \dots \int_0^\infty p_i(\boldsymbol{\Gamma}) f(\boldsymbol{\Gamma}) \, d\Gamma_1 \dots d\Gamma_{\mathrm{K}}, \tag{9}$$

and the identity between the expected values of the original Bernoulli variable $\mathbb{1}_i$ and the new Poisson variable Y_i is granted:

$$\mathbf{E}[Y_i] = \mathbf{E}[\mathbb{1}_i] = q_i. \tag{10}$$

According to the above hypothesis, the portfolio loss is now given by L_{Y}:

$$L_{\mathrm{Y}} = \sum_{i=1}^{N} Y_i \cdot E_i, \quad \text{where} \quad Y_i|\mathbf{\Gamma} \sim \text{Poisson}\left(p_i(\mathbf{\Gamma})\right). \tag{11}$$

In [1] it is shown how to compute the distribution of L_{Y} using a recursive method (for more detail on the so-called "Panjer recursion", see, *e.g.* [3]). The accuracy, stability and possible variants of the original algorithm are discussed in [2]. Numerical computation of the distribution by mean of Monte Carlo simulation is also possible and turns out to be particularly simple. Dedicated importance sampling algorithms are also available in the literature [4].

Notice that, although the distributions of L and L_{Y} differ, the *best estimate* of the portfolio loss is the same:

$$\mathbf{E}\left[L\right] = \mathbf{E}\left[L_{\mathrm{Y}}\right]. \tag{12}$$

In the language of copula functions, the structure of dependence implied by (7) corresponds [5] to a multivariate Clayton copula, *i.e.* an Archimedean copula where latent variables are gamma-distributed (for the relation between Archimedean copula functions and factor models see, *e.g*, [6, §2.1]). Notice that the parameters of the copula are the factor loadings ω_{ik}, and that they can be gathered, taking into account the normalization condition (8), in a $N \times K$ matrix W:

$$W = \begin{pmatrix} \omega_{11} & \dots & \omega_{1K} \\ \vdots & \ddots & \vdots \\ \omega_{N1} & \dots & \omega_{NK} \end{pmatrix}, \tag{13}$$

which is, for typical values of N and K, much smaller than the $N \times N$ covariance matrix between the default indicators $\mathbb{1}$.

It is easy [7] to show that :

$$\mathbf{cov}\left[Y_i, Y_j\right] = q_i\, q_j \sum_{k=1}^{K} \omega_{ik}\omega_{jk}\sigma_k^2 + \delta_{ij} q_i \sum_{k=0}^{K} \omega_{ik}. \tag{14}$$

As suggested in [7], equation (14) allows the calibration of the factor loadings, and thus of the dependence structure of the CreditRisk$^+$ model, by matching the observed covariance matrix of historical default time series with model values.

However, while the model has been defined with a reference "forecasting" time horizon $(t, T]$, that is typically of 1 year, *i.e.* $T = t + 1$, it is not *a priori* evident how to use historical time series with a different frequency. Naively, it is reasonable to expect that the higher the information provided by the historical time series (*i.e.* the higher the frequency) the better should be the calibration.

The following of this work will address this issue.

3 CreditRisk$^+$ using multiple unwind periods

In this section we discuss the internal consistency of the model hypothesis at different time scales. This allows us to write two estimators for the covariance matrix $A := W\Sigma W^T$, which is necessary in order to complete the calibration of the model. The estimators are defined using historical series sampled with a period that is not necessarily equal to the projection horizon that defines A.

3.1 The single unwind period case

In order to address the problem of defining CreditRisk$^+$ in a "roll-over" framework, defined by an arbitrary set of unwind periods, we introduce a (slighlty) modified definition of the relevant random variables in the single period case.

Notice that the r.v. Y_i introduced by the authors of CreditRisk$^+$ to model the default of the i-th risk can take values larger than 1, and – similarly – $p_i(\mathbf{\Gamma})$ can take values larger than 1. This is formally correct since Y_i is Poisson-distributed with parameter $p_i(\mathbf{\Gamma})$. The "Poisson approximation" introduced by substituting $\mathbb{1}_i$ with Y_i is numerically sound when:

$$q_i \ll \boldsymbol{\omega}_i \cdot \boldsymbol{\sigma}_\Gamma^2 \quad \wedge \quad q_i \ll 1, \quad i = 1, \ldots, N, \tag{15}$$

a condition that is well fulfilled in most relevant cases.

In order to recover the use of Bernoulli variables we modify the definition of Y_i as follows. Let $\widetilde{Y}_i(t,T)$ be a Bernoulli random variable that takes the value 1 if default of the i-th risk occurs in $(t,T]$ and 0 otherwise. Conditional on the market variables the distribution of $\widetilde{Y}_i(t,T)$ is then:

$$\widetilde{Y}_i(t,T) := \begin{cases} 1 & \tilde{p}_i(\boldsymbol{\omega}_i, \mathbf{\Gamma}, t, T), \\ 0 & 1 - \tilde{p}_i(\boldsymbol{\omega}_i, \mathbf{\Gamma}, t, T), \end{cases} \tag{16}$$

where:

$$\tilde{p}_i(\mathbf{w}_i, \mathbf{\Gamma}, t, T) = 1 - \exp\left[-q_i(t,T)\left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik}\,\Gamma_k\right)\right]. \tag{17}$$

With the above definition we have:

$$\mathbf{P}\left[Y_i = 0\right] = \mathbf{P}\left[\widetilde{Y}_i(t,T) = 0\right], \quad \mathbf{P}\left[Y_i \geq 1\right] = \mathbf{P}\left[\widetilde{Y}_i(t,T) = 1\right], \tag{18}$$

and

$$\mathbf{E}\left[Y_i\right] \simeq \mathbf{E}\left[\widetilde{Y}_i(t,T)\right], \tag{19}$$

where the approximation is due to the fulfillement of eq. (15). In particular, when $K = 1$ the expected value of $\widetilde{Y}_i(t,T)$ is:

$$\mathbf{E}\left[\widetilde{Y}_i(t,T)\right] = 1 - e^{-q\,\omega_{i0}}\left[1 + q\,(1-\omega_{i0})\sigma_1^2\right]^{-1/\sigma_1^2} = q + \mathcal{O}(q^2). \tag{20}$$

3.2 The multiple unwind periods case

Let $t \equiv t_0, t_1, \ldots, t_m \equiv T$ be a partition of the $(t, T]$ time interval. Supposing default events in each sub-interval are independent from each other, the following condition should hold:

$$1 - \tilde{p}_i(\boldsymbol{\omega}_i, \boldsymbol{\Gamma}, t, T) = \prod_{j=1}^{m} \left[1 - \tilde{p}_i(\boldsymbol{\omega}_i, \boldsymbol{\Gamma}^{(j)}, t_{j-1}, t_j)\right], \tag{21}$$

where $\boldsymbol{\Gamma}^{(j)}$ is the value of the latent variable $\boldsymbol{\Gamma}$ for the j-th sub-interval.

Since equation (17) has been introduced for any interval $(t, T]$, it should also hold for each sub-interval of the partition, *i.e.*:

$$p_i(\boldsymbol{\omega}_i, \boldsymbol{\Gamma}^{(j)}, t_{j-1}, t_j) = 1 - \exp\left[-q_i(t_{j-1}, t_j)\left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik}\Gamma_k^{(j)}\right)\right], \tag{22}$$

$$j = 1, \ldots, m, \quad i = 1, \ldots, N,$$

where it has been assumed, without loss of generality, that $\mathbf{E}[\Gamma_k^{(j)}] = \mathbf{E}[\Gamma_k] = 1$. Introducing (22) into (21) and taking the logarithm, it follows that:

$$\begin{aligned} q_i(t,T)\left(\omega_{i0} + \sum_{k=1}^{K}\omega_{ik}\Gamma_k\right) &= \sum_{j=1}^{m} q_i(t_{j-1}, t_j)\left(\omega_{i0} + \sum_{k=1}^{K}\omega_{ik}\Gamma_k^{(j)}\right) \\ \omega_{i0}\, q_i(t,T) + \sum_{k=1}^{K}\omega_{ik} q_i(t,T)\,\Gamma_k &= \omega_{i0}\sum_{j=1}^{m} q_i(t_{j-1}, t_j) + \sum_{k=1}^{K}\omega_{ik}\sum_{j=1}^{m} q_i(t_{j-1}, t_j)\,\Gamma_k^{(j)} \end{aligned} \tag{23}$$

Given that the sum of gamma-distributed random variables with the same scale parameter β are still gamma-distributed:

$$\sum_{\ell=1}^{n} \Gamma(\alpha_\ell, \beta) \sim \Gamma\left(\sum_{\ell=1}^{n} \alpha_\ell, \beta\right), \tag{24}$$

and that for any given positive constant c the following scaling property holds:

$$c\,\Gamma(\alpha, \beta) \sim \Gamma(\alpha, c\,\beta), \quad c > 0, \tag{25}$$

equation (23) can be satisfied by assuming that:

$$\text{Hyp.1} \qquad \frac{q_i\,(t,T)}{T-t} = \frac{q_i\,(t_{j-1}, t_j)}{t_j - t_{j-1}} = \text{constant} \quad j = 1, \ldots, m, \tag{26}$$

and

$$\text{Hyp.2} \quad \Gamma_k^{(j)} \sim \Gamma\left(\sigma_k^{-2}\,\frac{t_j - t_{j-1}}{T-t}, \sigma_k^2\,\frac{T-t}{t_j - t_{j-1}}\right) \quad j = 1, \ldots, m. \tag{27}$$

In this way the right term of eq. (23) reduces to:

$$\omega_{i0}\sum_{j=1}^{m}\underbrace{q_i(t,T)\frac{t_j - t_{j-1}}{T-t}}_{q_i(t_{j-1},t_j)} + \sum_{k=1}^{K}\omega_{ik}\sum_{j=1}^{m}\underbrace{q_i(t,T)\frac{t_j - t_{j-1}}{T-t}}_{q_i(t_{j-1},t_j)}\,\Gamma_k^{(j)}, \tag{28}$$

and further:

$$\omega_{i0}\, q_i(t,T) + \sum_{k=1}^{K} \omega_{ik}\, q_i(t,T) \underbrace{\sum_{j=1}^{m} \frac{t_j - t_{j-1}}{T-t}\, \Gamma_k^{(j)}}_{\sim\Gamma\left(\sigma_k^{-2},\sigma_k^2\right)} \tag{29}$$

that is equal to the left side of (23).

Remark. The first hypothesis can hold only approximately, since, in the limit of large time horizons $(t,T]$ it would violate the condition $q_i(t,T) \leq 1$. For typical values of $q_i(t,T)$ (*e.g.* 1%) and $(t,T]$ (*e.g.* one year) it can be considered a good approximation. The hypothesis can be restated in terms of constant default intensity λ_i, so that

$$q_i(t_{j-1},t_j) = 1 - e^{-\lambda_i(t_j - t_{j-1})} \simeq \lambda_i(t_j - t_{j-1}).$$

As an alternative, it is possible to modify eq. (17) by introducing a new variable $\widetilde{q}_i$:

$$\tilde{p}_i(\mathbf{w}_i, \mathbf{\Gamma}, t, T) = 1 - \exp\left[-\widetilde{q}_i(t,T)\left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik}\, \Gamma_k\right)\right]. \tag{30}$$

In this case $\widetilde{q}_i$ can take any (positive) value. ∎

Remark. The second hypothesis is fully consistent with the hypotheses made on the latent variables in the case of a single unwind period, *i.e.*

$$\mathbf{E}\left[\Gamma_k^{(j)}\right] = 1, \quad \mathbf{var}\left[\Gamma_k^{(j)}\right] = \sigma_k^2 \frac{T-t}{t_j - t_{j-1}} \xrightarrow[m\to 1]{} \sigma_k^2.$$

Notice that in case of a constant mesh of norm Δ_t:

$$(t_j - t_{j-1}) = \frac{T-t}{m} := \Delta_t, \quad j = 1, \dot{;} m,$$

the variance of $\Gamma_k^{(j)}$ scales with m:

$$\mathbf{var}\left[\Gamma_k^{(j)}\right] = m\, \sigma_k^2,$$

i.e. it is larger over smaller time intervals. ∎

4 Calibration of the structure of dependence

The calibration of the model is performed starting from a partition of the portfolio in M sets of homegeneous risks $c_h(t)$, $h = 1, \ldots, M$. The risks are homogeneous in the sense that the risks belonging to the same set $c_h(t)$ have the same vector of factor loadings $\boldsymbol{\omega}^{(h)}$. The sets have an explicit time dependence since they can change by the occurence of defaults. On the contrary, the structure of dependence, defined by $\boldsymbol{\omega}^{(h)}$ is supposed to be time independent.

4.1 The single unwind period case

The first case considered is that of a single unwind period (t, T). For each set $c_h(t)$, let $n_h(t) := |c_h(t)|$ be the number of risks in the set, F_h the default frequency, and G_h its complement to one. According to the standard CreditRisk$^+$ setting one has:

$$F_h := \frac{\sum_{i \in c_h(t)} Y_i}{n_h(t)}, \tag{31}$$

$$G_h := \frac{\sum_{i \in c_h(t)} (1 - Y_i)}{n_h(t)} = 1 - F_h. \tag{32}$$

The expected values of F_h and G_h are respectively:

$$\overline{q}_h := \mathbf{E}[F_h] = \frac{\sum_{i \in c_h(t)} q_i(t, T)}{n_h(t)}, \tag{33}$$

$$\overline{s}_h := \mathbf{E}[G_h] = 1 - \mathbf{E}[F_h]. \tag{34}$$

For any pair of sets of risks $\{h_1, h_2\}$, the covariance between the default frequencies is:

$$\begin{aligned} \mathbf{cov}(F_{h_1}, F_{h_2}) &= \mathbf{E}[(F_{h_1} - \mathbf{E}[F_{h_1}])(F_{h_2} - \mathbf{E}[F_{h_2}])] \\ &= \frac{1}{n_{h_1} n_{h_2}} \mathbf{E}\left[\sum_{i \in c_{h_1}} (Y_i - q_i) \sum_{i' \in c_{h_2}} (Y_{i'} - q_{i'})\right] \\ &= \frac{1}{n_{h_1} n_{h_2}} \sum_{i \in c_{h_1}} \sum_{i' \in c_{h_2}} \mathbf{cov}(Y_i, Y_{i'}), \end{aligned} \tag{35}$$

that, using eq. (14), becomes:

$$\mathbf{cov}(F_{h_1}, F_{h_2}) = \frac{1}{n_{h_1} n_{h_2}} \sum_{i \in c_{h_1}} \sum_{i' \in c_{h_2}} \left(q_i q_{i'} \sum_{k=1}^{K} \omega_{ik} \omega_{i'k} \sigma_k^2 + \delta_{ii'} q_i \right). \tag{36}$$

Eq. (36) shows the relation between the observed covariance of default frequencies and the factor loadings, describing the structure of dependence of the model.

Moreover, using the hypothesis that inside a homegenous set all risks have the same factor loadings, the above expression simplifies to:

$$\mathbf{cov}(F_{h_1}, F_{h_2}) = \overline{q}_{h_1} \overline{q}_{h_2} \sum_{k=1}^{K} \omega_{h_1 k} \omega_{h_2 k} \sigma_k^2 + \delta_{h_1 h_2} \frac{\overline{q}_{h_1}}{n_{h_1}} \tag{37}$$

Notice that the second term in eq. (37) is present only when $h_1 = h_2$, and becomes quickly negligible with increasing values of n_{h_1} (since $\overline{q}_{h_1} < 1$).

4.2 The multiple unwind period case

For the multiple unwind period case the analysis is restricted to time subintervals of equal length $\delta_t = (T-t)/m$. Similarly to the previous section, it is possibile to define the default frequencies or risks belonging to the set h, $F_h(t_{j-1}, t_j)$ $(j = 1, \ldots, m)$, and their complements to 1, $G_h(t_{j-1}, t_j)$, for each subinterval:

$$F_h^{(j)} := F_h(t_{j-1}, t_j), \tag{38}$$

$$G_h^{(j)} := 1 - F_h^{(j)}, \tag{39}$$

$$\overline{q}_h^{(j)} := \mathbf{E}\left[F_h^{(j)}\right], \tag{40}$$

$$\overline{s}_h^{(j)} := \mathbf{E}\left[G_h^{(j)}\right], \tag{41}$$

where:

$$t_j := t + j\frac{T-t}{m}, \quad j = 1, \ldots, m. \tag{42}$$

From eq. (38-41) it is possible to define two single observables:

$$F_{mh} := 1 - \prod_{j=1}^{m}\left[1 - F_h^{(j)}\right] \tag{43}$$

$$G_{mh} := \prod_{j=1}^{m} G_h^{(j)} = 1 - F_{mh} \tag{44}$$

For any pair of sets $\{h_1, h_2\}$ the covariance between F_{mh_1} and F_{mh_2} is given by:

$$\begin{aligned}
\mathbf{cov}\left(F_{mh_1}, F_{mh_2}\right) &= \mathbf{cov}\left(G_{mh_1}, G_{mh_2}\right) \\
&= \prod_{j=1}^{m} \mathbf{E}\left[G_{h_1}^{(j)} G_{h_2}^{(j)}\right] - \overline{s}_{h_1}\overline{s}_{h_2} \\
&= \prod_{j=1}^{m}\left[1 - \overline{q}_{h_1}^{(j)} - \overline{q}_{h_1}^{(j)} + \overline{q}_{h_1}^{(j)}\overline{q}_{h_1}^{(j)} + \mathbf{cov}\left(F_{h_1}^{(j)} F_{h_2}^{(j)}\right)\right] - \overline{s}_{h_1}\overline{s}_{h_2} \\
&= \prod_{j=1}^{m}\left[\mathbf{cov}\left(F_{h_1}^{(j)} F_{h_2}^{(j)}\right) + \overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}\right] - \overline{s}_{h_1}\overline{s}_{h_2}
\end{aligned} \tag{45}$$

Since all subintervals have the same length, the frequencies $F_h^{(j)}$ are *i.i.d.*, so that the above expression simplifies to:

$$\mathbf{cov}\left(F_{mh_1}, F_{mh_2}\right) + \overline{s}_{h_1}\overline{s}_{h_2} = \left[\mathbf{cov}\left(F_{h_1}^{(j)}, F_{h_2}^{(j)}\right) + \overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}\right]^m \tag{46}$$

for any $j = 1, \ldots, m$.

In the limit of "large portfolio", *i.e.* $n_h(t_j) \to \infty$ $(j = 1, \ldots, m)$, from eqs. (7) and (21) one has:

$$\lim_{n_h(t_1), \ldots, n_h(t_m) \to \infty} F_{mh} = \lim_{n_h(t_1) \to \infty} F_h \simeq \overline{q}_h\left(\omega_{h0} + \sum_{k=1}^{K} \omega_{hk}\Gamma_k\right), \tag{47}$$

therefore both $F_{mh}(t,T)$ and $F_h(t,T)$ are estimators of the default frequency for the $(t,T]$ interval. Thus, eq. (46) can be rewritten as:

$$\mathbf{cov}\left(F_{h_1},F_{h_2}\right)+\overline{s}_{h_1}\overline{s}_{h_2}=\left[\mathbf{cov}\left(F_{h_1}^{(j)},F_{h_2}^{(j)}\right)+\overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}\right]^m \tag{48}$$

and, since $m=(T-t)/(t_2-t_1)$,

$$\left[\mathbf{cov}\left(F_{h_1},F_{h_2}\right)+\overline{s}_{h_1}\overline{s}_{h_2}\right]^{1/(T-t)}=\left[\mathbf{cov}\left(F_{h_1}^{(j)},F_{h_2}^{(j)}\right)+\overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}\right]^{1/(t_2-t_1)}, \tag{49}$$

so that, in general, the following equation holds:

$$\left[\mathbf{cov}\left(F_{h_1}\left(t,t'\right),F_{h_2}\left(t,t'\right)\right)+\overline{s}_{h_1}\left(t,t'\right)\overline{s}_{h_2}\left(t,t'\right)\right]^{\frac{1}{t-t'}}=\text{constant}. \tag{50}$$

Eq. (50), together with eq. (57) and (68), is one of the main results of this work. It allows to build and estimator of $\mathbf{cov}\left(F_{h_1}\left(t,T\right),F_{h_2}\left(t,T\right)\right)$ using default frequencies $F_h\left(t_1,t_2\right)$ defined on a different time interval. The dependence upon m of the precision of the covariance estimator will be discussed in section 5.

Finally, inserting $\mathbf{cov}\left(F_{mh_1},F_{mh_2}\right)$ into eq. (37) allows to perform the calibration of the structure of dependence of the CreditRisk$^+$ model, by first determining the elements of the A matrix, whose elements are defined as:

$$\begin{aligned}
A_{h_1h_2} &:= \sum_{k=1}^{K} w_{1k}w_{2k}\,\sigma_k^2, \qquad (51)\\
&= \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\left[\mathbf{cov}\left(F_{mh_1},F_{mh_2}\right)-\delta_{h_1h_2}\frac{\overline{q}_{h_1}}{n_{h_1}}\right]\\
&= \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\left[\left(\mathbf{cov}\left(F_{h_1}^{(j)},F_{h_2}^{(j)}\right)+\overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}\right)^m-\overline{s}_{h_1}\overline{s}_{h_2}-\delta_{h_1h_2}\frac{\overline{q}_{h_1}}{n_{h_1}},\right]
\end{aligned}$$

for any $j=1,\ldots,m$, and then decomposing A, thus obtaning a separate estimate of the $\left\{W,\boldsymbol{\sigma}_\Gamma^2\right\}$ parameters. The decompostion can be performed, *e.g.*, by using the technique described in [7].

4.3 The exponential case

In this paragraph the problem of calibrating the structure of dependence will be addressed using the exponential form of the conditional default frequencies introduced in equation (17), which has been proven to be consistent when considering multiple unwind periods in section 3.2. Since now $\widetilde{Y}_i$ variables are used instead of the corresponding variables Y_i, the frequencies F_h and their complements G_h are replaced by $\widetilde{F}_h$ and $\widetilde{G}_h$, defined in a similar fashion.

Given a generic time interval (t_1,t_2), equations (16), (17) and (32) imply that

$$\lim_{n_h(t_1)\to\infty}-\ln\widetilde{G}_h(t_1,t_2)=\left[\omega_{h0}+\sum_{k=1}^{K}\omega_{hk}\Gamma_k(t_1,t_2)\right]q_h^\star(t_1,t_2), \tag{52}$$

where

$$q_h^\star(t_1,t_2):=-\ln\frac{\sum_{j\in c_h(t_1)}e^{-q_j(t_1,t_2)}}{n_h(t_1)}. \tag{53}$$

In general $\overline{q}_h(t_1,t_2) \neq q^\star_h(t_1,t_2)$, unless $q_i = \overline{q}_h$ for each risk $i \in c_h$. However, in realistic cases, the whole set $\{q_j | j \in c_h\}$ should be available from the scoring/pricing model of the financial institution which is calibrating the CreditRisk$^+$ model.
Equations (21) and (52) allow for computing the exact version of equation (47):

$$\lim_{n_h(t_1),\ldots,n_h(t_m)\to\infty} \sum_{i=1}^{m} L_h^{(i)} = \lim_{n_h(t_1)\to\infty} L_h = \overline{q}_h \left(\omega_{h0} + \sum_{k=1}^{K} \omega_{hk}\Gamma_k \right) \tag{54}$$

where L_h, $L_h^{(i)}$ $(i = 1,\ldots,m)$ are defined as follows

$$\begin{aligned} L_h &:= -\frac{\overline{q}_h(t,T)}{q^\star_h(t,T)} \ln \widetilde{G}_h(t,T) \\ L_h^{(i)} &:= -\frac{\overline{q}_h(t_{i-1},t_i)}{q^\star_h(t_{i-1},t_i)} \ln \widetilde{G}_h(t_{i-1},t_i) \end{aligned}$$

with the same notation used in equation (46). Definitions above imply directly that

$$\mathbf{cov}\left[L_{h_1}^{(i)}, L_{h_2}^{(i')}\right] = \delta_{ii'}\mathbf{cov}\left[L_{h_1}^{(i)}, L_{h_2}^{(i)}\right] \tag{55}$$

and together with equation (26) allow for the exponential case version of equations (46, 50)

$$\forall i \in \{1\ldots m\} \quad \mathbf{cov}\left[L_{h_1}, L_{h_2}\right] = m\ \mathbf{cov}\left[L_{h_1}^{(i)}, L_{h_2}^{(i)}\right] \tag{56}$$

$$\frac{1}{t_2 - t_1}\mathbf{cov}\left[L_{h_1}(t_1,t_2), L_{h_2}(t_1,t_2)\right] = c' \tag{57}$$

where c' is a generic constant. Here we are supposing that the portfolio composition is stable over time, implying that $q^\star_{h_1}(t_{i-1},t_i)$ does not depend on i.

Hence equation (52) in the exponential case becomes

$$\begin{aligned} A_{h_1h_2}(t,T) &= \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\mathbf{cov}\left[L_{h_1}, L_{h_2}\right] \\ &= \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}} \sum_{i=1}^{m} \mathbf{cov}\left[\frac{\overline{q}_{h_1}(t_{i-1},t_i)}{q^\star_{h_1}(t_{i-1},t_i)} \ln\left(1 - \widetilde{F}_{h_1}^{(i)}\right), \frac{\overline{q}_{h_2}(t_{i-1},t_i)}{q^\star_{h_2}(t_{i-1},t_i)} \ln\left(1 - \widetilde{F}_{h_2}^{(i)}\right)\right] \end{aligned} \tag{58}$$

where we have neglected the contribution of $\mathbf{cov}(\widetilde{Y}_j, \widetilde{Y}_j) \propto \frac{1}{n_h(t_1)} \simeq 0$.
Equation (58) holds also in the general case of a credit risk portfolio whose composition changes over time, leading to an actual time dependence of $\tilde{q}_h^{(i)}$ on i. In case of constant portfolio composition, $L_h^{(i)}$ $(i = 1\ldots m)$ are identically distributed and equation (58) simplifies to

$$A_{h_1h_2}(t,T) = \frac{1}{m}\frac{1}{q^\star_{h_1}(t_{i-1},t_i)q^\star_{h_2}(t_{i-1},t_i)}\mathbf{cov}\left[\ln\left(1 - \widetilde{F}_{h_1}^{(i)}\right), \ln\left(1 - \widetilde{F}_{h_2}^{(i)}\right)\right] \tag{59}$$

5 On the advantage of high frequency sampling

Suppose to have a (future) time horizon of interest of length Δt and a set of historical time series of defaults that span a (past) time interval of length $n\Delta t$. Typical examples can be $\Delta t = 1$ year and $n \in [10, 20]$. Moreover, suppose that the historical time series can be sampled with a period δt, which is m times smaller than Δt, $\delta t = \Delta t/m$, *e.g.* for $\Delta t = 1$ year m can typically be $m = 4$ (quarterly time series) or $m = 12$ (monthly time series) . Therefore, the historical time series are defined over $m \times n$ intervals of length δt, defined by a tenor $t_0, \ldots, t_{m\times n}$.

The question addressed in this section is if there is any advantage in using in calibration historical default time series with a period smaller that the time horizon on which the model is used to measure risk, also when statistical errors are taken into account. It will be shown that the statistical error on the determination of A using the observed time series of default frequencies depends on m, *i.e.* on the sampling frequency of the observations. Moreover, under the hypotheses reported in (64) and (60), the statistical error can be determined in closed form as function of m.

As in the previous paragraph, the case of the standard CreditRisk$^+$ hypothesis of the conditional default frequencies (linear in the latent variables) and the modified hypothesis introduced in section 3.1 (exponential in the latent variables) will be discussed. In addition, paragraph 5.4 will cope with the case where the hypothesis in (64) does not hold, addressing the problem in a numerical way.

The first simplifying hypothesis is the hypothesis of "large" portfolio:

$$n_h(t_k) \gg 1 \tag{60}$$

for any time t_k, $k = 1, \ldots, m \times n$, and any set h,. Under this hypothesis one has that:

$$\begin{gathered} V\left[G_h(t_k)|\mathbf{\Gamma}(t_k)\right] \simeq 0; \quad V\left[L_h(t_k)|\mathbf{\Gamma}(t_k)\right] \simeq 0; \quad V\left[\overline{s}_h(t_k)|\mathbf{\Gamma}(t_k)\right] \simeq 0; \\ V\left[\overline{q}_h(t_k)|\mathbf{\Gamma}(t_k)\right] \simeq 0; \quad V\left[\tilde{q}_h(t_k)|\mathbf{\Gamma}(t_k)\right] \simeq 0, \end{gathered} \tag{61}$$

and all the uncertainty is due to the randomness of the latent variables.

The second simplifying hypothesis is the hypothesis of "quasi-gaussian regime", that allows to use the property of the Wishart distribution. In particular, for any couple of random variables $X_1(t)$ and $X_2(t)$, defined over the tenor times $t_k = t_0 + k\,\delta t$, $k = 1, \ldots, m \times n$, the estimator of the covariance $\mathbf{cov}[X_1, X_2]$ is:

$$\begin{aligned} \widehat{V}_{12} &:= \frac{1}{m \cdot n - 1} \sum_{k=1}^{m\cdot n} \left(X_1(t_k) - \widehat{\mu}_1\right)\left(X_2(t_k) - \widehat{\mu}_2\right), \\ \widehat{\mu}_j &:= \frac{1}{m \cdot n} \sum_{k=1}^{m\cdot n} X_j(t_k) \quad j = 1, 2. \end{aligned}$$

When X_1 and X_2 are jointly normal with covariance matrix V, then the covariance estimator is distributed as a Wishart random variable with $m \times n - 1$ degrees of freedom, with expected value and variance respectively given by:

$$\mathbf{E}\left[\widehat{V}_{12}\right] = V_{12}, \tag{62}$$

$$\mathbf{var}\left[\widehat{V}_{12}\right] = \frac{1}{m \cdot n - 1}\left(V_{12}^2 + V_{11}V_{22}\right). \tag{63}$$

In section 5.3 it will be shown that under the hypothesis of "small" variances of the latent factors:

$$\sigma_k^2 \ll 1, \quad k = 1, \ldots, K, \tag{64}$$

the gamma distributions can be well approximated with normal distributions, so that the property of the Wishart distribution can be used.

Under the above simplifying hypotheses the estimators of $A_{h_1 h_2}$, respectively corresponding to eq. (52), for the linear case, and to eq. (59) for the exponential case are then:

$$\hat{A}_{h_1 h_2}^{(L,m)} := \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}} \left(\mathbf{c\hat{o}v}_m(F_{h_1}, F_{h_1}) + [\overline{s}_{h_1}\overline{s}_{h_2}]^{\frac{1}{m}} \right)^m - \overline{s}_{h_1}\overline{s}_{h_2}, \tag{65}$$

$$\hat{A}_{h_1 h_2}^{(E,m)} := \frac{1}{m}\frac{1}{\tilde{q}_{h_1}\tilde{q}_{h_2}} \mathbf{c\hat{o}v}_m(L_{h_1}, L_{h_1}). \tag{66}$$

In both cases, the improvement in statistical precision with respect to the estimate with no subsampling, will be measured by the following ratio:

$$\varepsilon\left[\hat{A}_{h_1 h_2}^{(m)}\right] := \frac{\sqrt{\mathbf{var}\left[\hat{A}_{h_1 h_2}^{(m)}\right]}}{\sqrt{\mathbf{var}\left[\hat{A}_{h_1 h_2}^{(1)}\right]}}. \tag{67}$$

In sections 5.1 and 5.2 it will be shown that in the quasi-Gaussian regime the precison $\varepsilon\left[\hat{A}_{h_1 h_2}^{(m)}\right]$ is given, respectively for the linear and exponential case, by:

$$\varepsilon\left[\hat{A}_{h_1 h_2}^{(L,m)}\right] \simeq \varepsilon\left[\hat{A}_{h_1 h_2}^{(E,m)}\right] \simeq \sqrt{\frac{n-1}{m \cdot n - 1}} \xrightarrow[n\to\infty]{} \frac{1}{\sqrt{m}}. \tag{68}$$

Hence, for a large portfolio, in the quasi-gaussian regime, it is statistically convienent to sample the historical time series as frequently as possible.

5.1 The linear case

In this paragraph the expressions reported in eqs. (65) and (68) are derived. Let us introduce the notation:

$$c_{h_1h_2}^{(j)} := \mathbf{cov}\left(F_{h_1}^{(j)}, F_{h_2}^{(j)}\right) + \overline{s}_{h_1}^{(j)}\overline{s}_{h_2}^{(j)}. \tag{69}$$

The variables $\left\{c_{h_1h_2}^{(j)}\right\}_{j=1,\ldots,m}$ are independent and identically distributed. Furthermore, equation (52) implies that:

$$A_{h_1h_2}^{(L,m)} = \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\left(\prod_{j=1}^{m} c_{h_1h_2}^{(j)} - \overline{s}_{h_1}\overline{s}_{h_2}\right). \tag{70}$$

Moreover hypothesis (60) implies:

$$\begin{aligned}\sqrt{\mathbf{var}\left[A_{h_1h_2}^{(L,m)}\right]} &= \sqrt{\mathbf{var}\left[\frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\prod_{j=1}^{m} c_{h_1h_2}^{(j)} - \overline{s}_{h_1}\overline{s}_{h_2}\right]} \\ &\simeq \frac{1}{\overline{q}_{h_1}\overline{q}_{h_2}}\sqrt{\mathbf{var}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right]} \end{aligned} \tag{71}$$

given that $c_{h_1h_2}^{(j)}$ $(j = 1\ldots m)$ are independent, it follows that:

$$\begin{aligned}\mathbf{var}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right] &= \mathbf{E}\left[\left(\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right)^2\right] - \left(\mathbf{E}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right]\right)^2 \\ &= \prod_{j=1}^{m}\mathbf{E}\left[\left(c_{h_1h_2}^{(j)}\right)^2\right] - \prod_{j=1}^{m}\mathbf{E}\left(\left[c_{h_1h_2}^{(j)}\right]\right)^2, \end{aligned} \tag{72}$$

and considering that $c_{h_1h_2}^{(j)}$ $(j = 1\ldots m)$ are identically distributed:

$$\begin{aligned}\mathbf{var}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right] &= \left(\mathbf{E}\left[\left(c_{h_1h_2}^{(1)}\right)^2\right]\right)^m - \mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{2m} \\ &= \left(\mathbf{var}\left[c_{h_1h_2}^{(1)}\right] - \mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^2\right)^m - \mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{2m} \end{aligned} \tag{73}$$

Equations (50) and (63) imply that:

$$\mathbf{E}\left[c_{h_1h_2}^{(1)}\right] = \left(\overline{q}_{h_1}\overline{q}_{h_2}A_{h_1h_2}^{(L,m)} + \overline{s}_{h_1}\overline{s}_{h_2}\right)^{\frac{1}{m}} \tag{74}$$

$$\begin{aligned}\mathbf{var}\left[c_{h_1h_2}^{(1)}\right] &\simeq \mathbf{var}\left[\mathbf{cov}\left(F_{h_1}^{(i)}, F_{h_2}^{(i)}\right)\right] \\ &= \frac{1}{m\cdot n-1}\left[\mathbf{cov}^2\left(F_{h_1}^{(1)}, F_{h_2}^{(1)}\right)\mathbf{cov}\left(F_{h_1}^{(1)}, F_{h_1}^{(1)}\right)\mathbf{cov}\left(F_{h_2}^{(1)}, F_{h_2}^{(1)}\right)\right] \\ &= \left(\rho_{h_1h_2}^2 + 1\right)\left(\sigma_{h_1}^{(1)}\sigma_{h_2}^{(1)}\right)^2 \end{aligned} \tag{75}$$

Hence it follows from equation (73) and equation (75) that:

$$\mathbf{var}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right] = \sum_{j=1}^{m-1} (m{\cdot}n{-}1)^{-(m-j)} \left[\left(\rho_{h_1h_2}^2+1\right)\left(\sigma_{h_1}^{(1)}\sigma_{h_2}^{(1)}\right)^2\right]^{m-j} \mathbf{E}\left(\left[c_{12}^{(j)}\right]\right)^{2j} \tag{76}$$

that, together with the hypothesis in equation (64), implies that the series can be approximated by its leading term:

$$\mathbf{var}\left[\prod_{j=1}^{m} c_{h_1h_2}^{(j)}\right] \simeq (m\cdot n-1)^{-1}\left(\rho_{h_1h_2}^2+1\right)\left(\sigma_{h_1}^{(1)}\sigma_{h_2}^{(1)}\right)^2 \mathbf{E}\left(\left[c_{12}^{(j)}\right]\right)^{2(m-1)}. \tag{77}$$

This implies that:

$$\begin{aligned}
\varepsilon\left[A_{h_1h_2}^{(L,m)}\right] &\simeq \sqrt{\frac{(m\cdot n-1)^{-1}\left(\rho_{h_1h_2}^2+1\right)\left(\sigma_{h_1}^{(1)}\sigma_{h_2}^{(1)}\right)^2 \mathbf{E}\left(\left[c_{12}^{(i)}\right]\right)^{2(m-1)}}{(n-1)^{-1}\left(\rho_{h_1h_2}^2+1\right)\left(\sigma_{h_1}\sigma_{h_2}\right)^2}} \\
&= \sqrt{\frac{n-1}{m\cdot n-1}}\frac{\left(\rho_{h_1h_2}^2+1\right)^{1/2}\sigma_{h_1}^{(1)}\sigma_{h_2}^{(1)}\mathbf{E}\left(\left[c_{h_1h_2}^{(i)}\right]\right)^{(m-1)}}{\left(\rho_{h_1h_2}^2+1\right)^{1/2}\sigma_{h_1}\sigma_{h_2}}.
\end{aligned}$$

Multiplying and dividing by $\rho_{h_1h_2}/\left(\rho_{h_1h_2}^2+1\right)^{1/2}$ gives:

$$\begin{aligned}
\varepsilon\left[A_{h_1h_2}^{(L,m)}\right] &\simeq \sqrt{\frac{N-1}{m\cdot n-1}}\frac{\mathbf{cov}\left(F_{h_2}^{(1)},F_{h_2}^{(1)}\right)\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1}}{\mathbf{cov}\left(F_{h_2},F_{h_2}\right)} \\
&= \sqrt{\frac{n-1}{m\cdot n-1}}\frac{\left[\mathbf{cov}\left(F_{h_2}^{(1)},F_{h_2}^{(1)}\right)+\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)}-\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)}\right]\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1}}{\mathbf{cov}\left(F_{h_2},F_{h_2}\right)+\overline{s}_{h_1}\overline{s}_{h_2}-\overline{s}_{h_1}\overline{s}_{h_2}} \\
&= \sqrt{\frac{n-1}{m\cdot n-1}}\frac{\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m}-\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)}\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1}}{E\left[c_{h_1h_2}\right]-\overline{s}_{h_1}\overline{s}_{h_2}} \\
&= \sqrt{\frac{n-1}{m\cdot n-1}}\frac{\mathbf{E}\left(\left[c_{h_1h_2}\right]\right)-\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)}\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1}}{\mathbf{E}\left(\left[c_{h_1h_2}\right]\right)-\overline{s}_{h_1}\overline{s}_{h_2}}
\end{aligned}$$

Hypothesis in equation (64) implies that:

$$\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1} \simeq (\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)})^{m-1}, \tag{78}$$

and hence it follows that:

$$\frac{\mathbf{E}\left(\left[c_{h_1h_2}\right]\right)-\overline{s}_{h_1}^{(1)}\overline{s}_{h_2}^{(1)}\mathbf{E}\left(\left[c_{h_1h_2}^{(1)}\right]\right)^{m-1}}{\mathbf{E}\left(\left[c_{h_1h_2}\right]\right)-\overline{s}_{h_1}\overline{s}_{h_2}} \simeq 1, \tag{79}$$

which implies the result in equation (68).

5.2 The exponential case

In this paragraph, the expression reported in equation (68) are derived. On the other hand, the result in equation (66) is not discussed, since it follows from equation (59) directly.

From hypothesis in equation (60), it follows that:

$$\begin{aligned}
\sqrt{\mathbf{var}\left[\hat{A}^{(E,m)}_{h_1h_2}\right](t,T)} &= \sqrt{\mathbf{var}\left[\frac{1}{m}\frac{1}{\tilde{q}^{(j)}_{h_1}\tilde{q}^{(j)}_{h_2}}\mathbf{c\hat{o}v}_m\left[\ln\left(1-F^{(j)}_{h_1}\right),\ln\left(1-F^{(j)}_{h_2}\right)\right]\right]} \\
&\simeq \frac{1}{m}\frac{1}{\tilde{q}^{(j)}_{h_1}\tilde{q}^{(j)}_{h_2}}\sqrt{\mathbf{var}\left[\mathbf{c\hat{o}v}_m\left[\ln\left(1-F^{(j)}_{h_1}\right),\ln\left(1-F^{(j)}_{h_2}\right)\right]\right]} \\
&= \frac{1}{m}\frac{1}{\tilde{q}^{(j)}_{h_1}\tilde{q}^{(j)}_{h_2}}\sqrt{\frac{m^2\left(\tilde{q}^{(j)}_{h_1}\tilde{q}^{(j)}_{h_2}\right)^2}{m\cdot n-1}\left(A^2_{h_1h_2}+A_{h_1h_1}A_{h_2h_2}\right)} \\
&= \sqrt{\frac{1}{m\cdot n-1}\left(A^2_{h_1h_2}+A_{h_1h_1}A_{h_2h_2}\right)}, \qquad (80)
\end{aligned}$$

leading to the result in equation (68) for the exponential case.

5.3 Gaussian approximation for the latent variables

As previously discussed in paragraph 3.2, the latent variables $\Gamma^{(j)}_k$ $(k=1,\dots,K)$ are gamma-distributed random variables:

$$\Gamma^{(j)}_k \sim \Gamma\left(\theta_k^{-1},\theta_k\right), \quad \mathbf{E}\left[\Gamma^{(j)}_k\right]=1, \quad \sigma_k^2=\theta_k, \quad j=1,\dots,m.$$

Hence their probability densities $dF_k(x)$ satisfy the following:

$$dF_k(x) \propto x^{\theta_k^{-1}-1}\exp\left(-\theta_k^{-1}x\right)dx \qquad (81)$$

Under hypothesis in equation (64) it holds $\theta^{-1}-1\simeq\theta^{-1}$; thus it is approximately true that:

$$dF_k(x) \propto \exp\left(\frac{\ln x - x}{\theta_k}\right)dx. \qquad (82)$$

By introducing the auxiliary variable $x' := x-1$, it follows that:

$$\begin{aligned}
dF_k\left(x(x')\right) &\propto \exp\left(\frac{\ln(1+x')-x'-1}{\theta_k}\right)dx' \\
&\simeq \exp\left(-\frac{x'^2/2-1}{\theta_k}\right)dx' \\
&\propto \exp\left(-\frac{x'^2}{2\sigma_k^2}\right)dx'
\end{aligned}$$

where $\ln(1+x')$ has been approximated with the first three terms of its Maclaurin series. Centering the series in $x'=0$ (*i.e.* in $x=1$, the mean of each $\Gamma^{(i)}_k$) makes

the approximation more precise in the region of x where the probability density is higher. In the limit $\sigma_k^2 \to 0$ it holds that:

$$\Gamma_k^{(i)} \sim \mathcal{N}\left(\mu = 1, \sigma^2 = \theta_k = \sigma_k^2\right), \tag{83}$$

allowing to consider the covariance estimators approximately distributed as Wishart variables.

5.4 Beyond the quasi-gaussian regime: numerical simulations

In this section we verify that both the estimators $\hat{A}^{(E,m)}_{h_1h_2}$ and $\hat{A}^{(L,m)}_{h_1h_2}$ are more precise at increasing m. The analitycal results obtained in the quasi-gaussian regime, discussed in paragraph 5.3, are satisfied when the factor volatilities $\boldsymbol{\sigma}_\Gamma$ are much less than 1. At increasing σ_k $(k = 1, \ldots, K)$the distributions of the estimators change and the difference of precision amongst various m values becomes smaller. However, the error of $\hat{A}^{(m)}_{h_1h_2}$ results to be decreasing at increasing m even far from the quasi-gaussian regime.
Our case study is a two-factors-world (Γ_k, $k = 1, 2$). The dynamic of the couple of systemic factors induces the dependence between two populations of risks, as per the weights reported in Table 1.

k	0	1	2
w_{1k}	0,30	0,40	0,30
w_{2k}	0,50	0,25	0,25

Table 1: Matrix of weights used for the numerical simulations.

The volatilities (σ_k, $k = 1, 2$) associated to the factors are chosen as

$$\boldsymbol{\sigma}_\Gamma := 2^{i_\sigma} \begin{pmatrix} 2.5 \cdot 10^{-2} \\ 5.0 \cdot 10^{-2} \end{pmatrix}, \quad i_\sigma = 0 \ldots 6$$

We simulated the distributions of $\hat{A}^{(E,m)}_{12}$ and $\hat{A}^{(L,m)}_{12}$ $(m = 1 \ldots 12)$ using 10^5 scenarios of $\{F_1(t, n_1), F_2(t, n_2)\}$ where $t \in (t_0, t_0 + n\Delta T]$ $(n = 10)$ and n_h $(h = 1, 2)$ is the number of risks belonging to each cluster. In both cases we simulated the $\mathbf{F}_i(t, n_i)$ dynamic using equation (17). Each considered risk in each cluster is supposed to have the same unconditioned intensity of default

$$q_i(t, t + \Delta t) = -\frac{1}{\Delta t} \log(0.99)$$

The numerical result reported below have been produced in unit $\Delta t \equiv 1$. n_h allows us to take into account an additional contribution to the error $\sigma\left[\hat{A}_{12}\right]$, generated by the finiteness of each considered cluster. To this purpose we extract the number of claims per each elementary temporal step $\Delta t = 1/m$ from a binomial distribution with parameter $n_h \in \left\{10^3, 2.5 \cdot 10^3, 5 \cdot 10^3, 10^4, 2.5 \cdot 10^4, 5 \cdot 10^4\right\}$, for each cluster h considered. The underlying *ansatz* is that each defaulted risk is instantly replaced by a new risk, keeping each n_h constant in time. The case $n_h = \infty$ (absence of binomial source of randomness) is also considered.

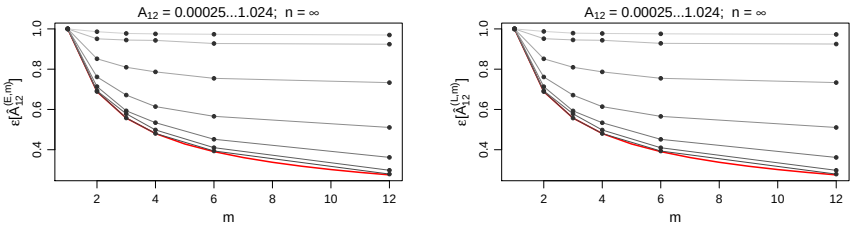

Figure 1: $\varepsilon\left[\hat{A}_{12}^{(E,m)}\right]$ (left) compared with $\varepsilon\left[\hat{A}_{12}^{(L,m)}\right]$ (right) as a function of m, considering increasing i_σ (from darker to brighter curve). This case excludes the binomial error ($n_h = \infty$). The red curve is the theoretical value of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ as a function of m in the quasi-gaussian regime.

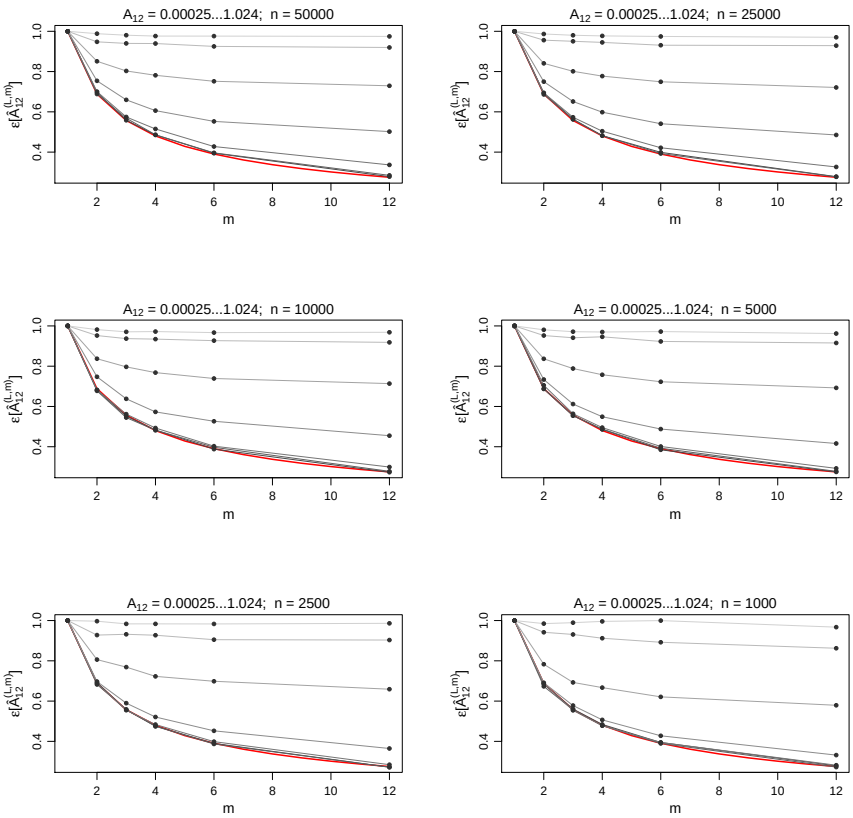

Figure 2: $\varepsilon\left[\hat{A}_{12}^{(L,m)}\right]$ as a function of m, considering increasing i_σ (from darker to brighter curve) and various choices of n_h. The red curve is the theoretical value of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ as a function of m in the quasi-gaussian regime. For $\sigma_1, \sigma_2 \ll 1$ the analytical result is perfectly satisfied. However, $\varepsilon\left[\hat{A}_{12}^{(L,m)}\right]$ is shown to be a decreasing function of m in general. Comparing this result with the $n_h = \infty$ case, we can state that $\varepsilon\left[\hat{A}_{h_1 h_2}^{(L,m)}\right]$ is almost insensitive to n_h, even for $n_h = 10^3$ (the minimum tested value).

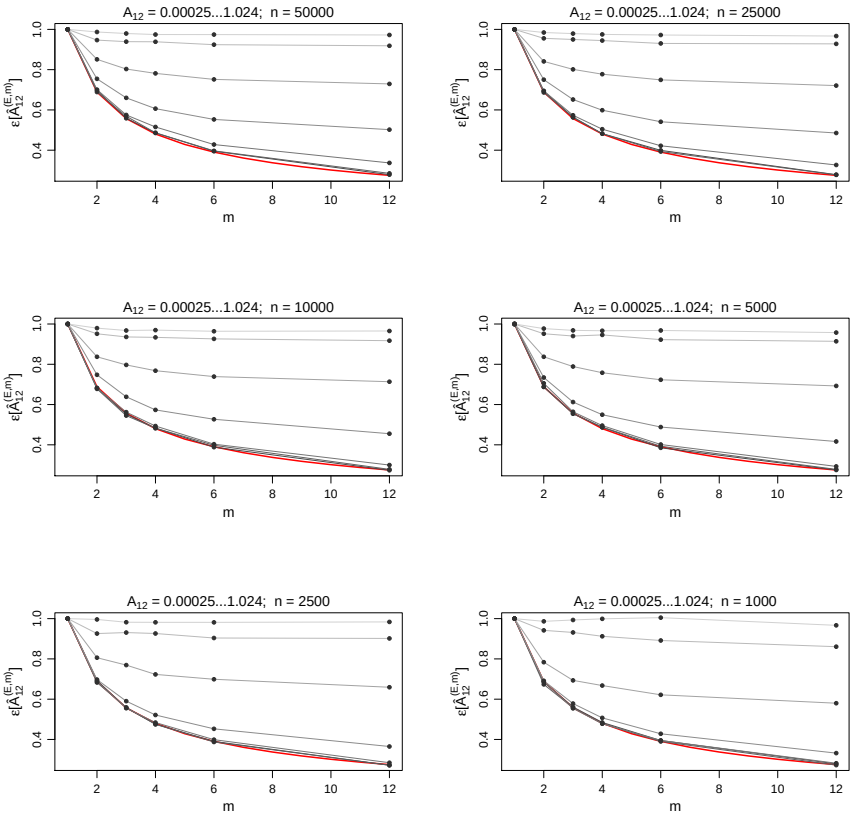

Figure 3: $\varepsilon\left[\hat{A}_{12}^{(E,m)}\right]$ as a function of m, considering increasing i_σ (from darker to brighter curve) and various choices of n_h. The red curve is the theoretical value $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ as a function of m in the quasi-gaussian regime. For $\sigma_1, \sigma_2 \ll 1$, the analytical result is perfectly satisfied. However $\varepsilon\left[\hat{A}_{12}^{(E,m)}\right]$ is shown to be a decreasing function of m in general. Comparing this result with the $n = \infty$ case, we can state that $\varepsilon\left[\hat{A}_{h_1 h_2}^{(E,m)}\right]$ is almost insensitive to n_h, even for $n_h = 10^3$ (the minimum tested value).

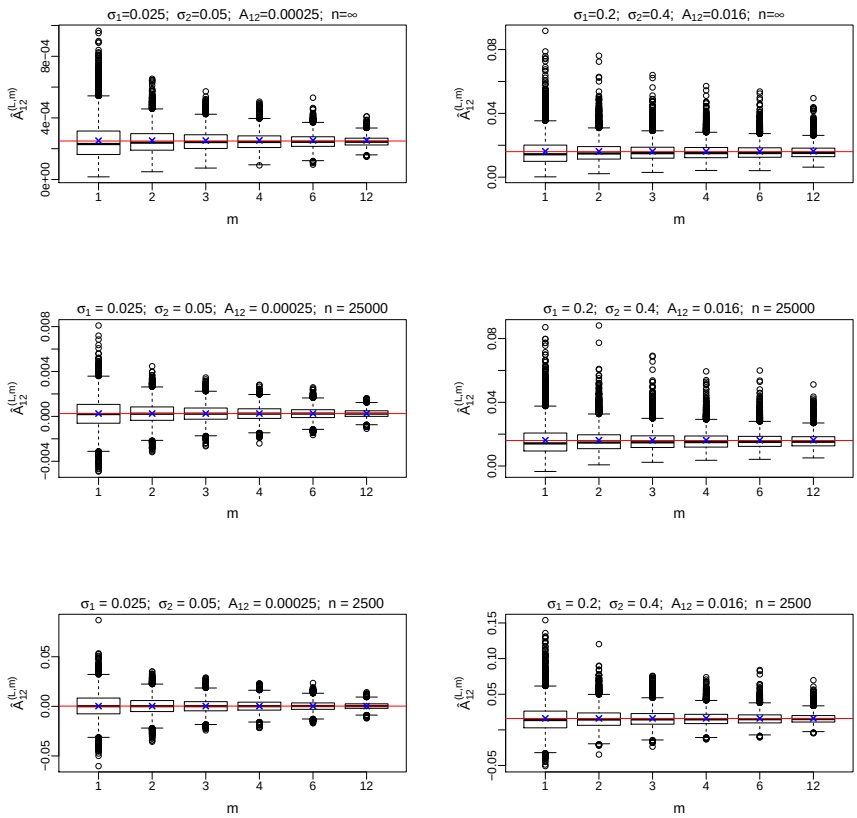

Figure 4: Boxplot of $\hat{A}_{12}^{(L,m)}$ as a function of m, considering three choices for n_h combined with two choices for $\boldsymbol{\sigma}_\Gamma$. The red line represent the true value of A_{12} and the blue X's stand for the average value of $\hat{A}_{12}^{(L,m)}$.

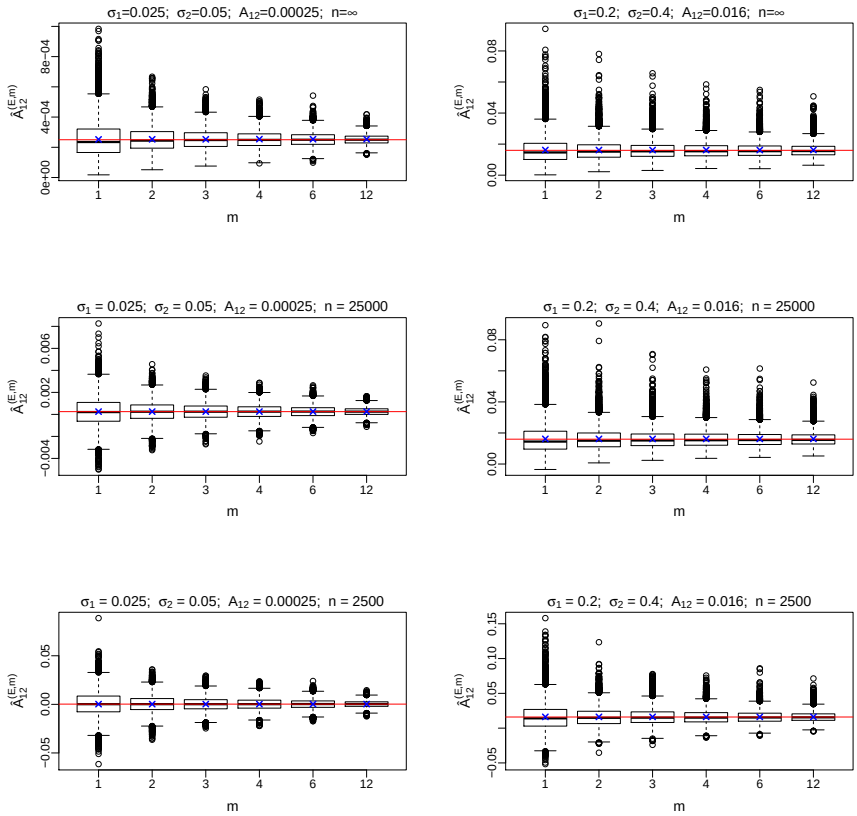

Figure 5: Boxplot of $\hat{A}_{12}^{(E,m)}$ as a function of m, considering three choices for n_h combined with two choices for $\boldsymbol{\sigma}_\Gamma$. The red line represent the true value of A_{12} and the blue X's stand for the average value of $\hat{A}_{12}^{(E,m)}$.

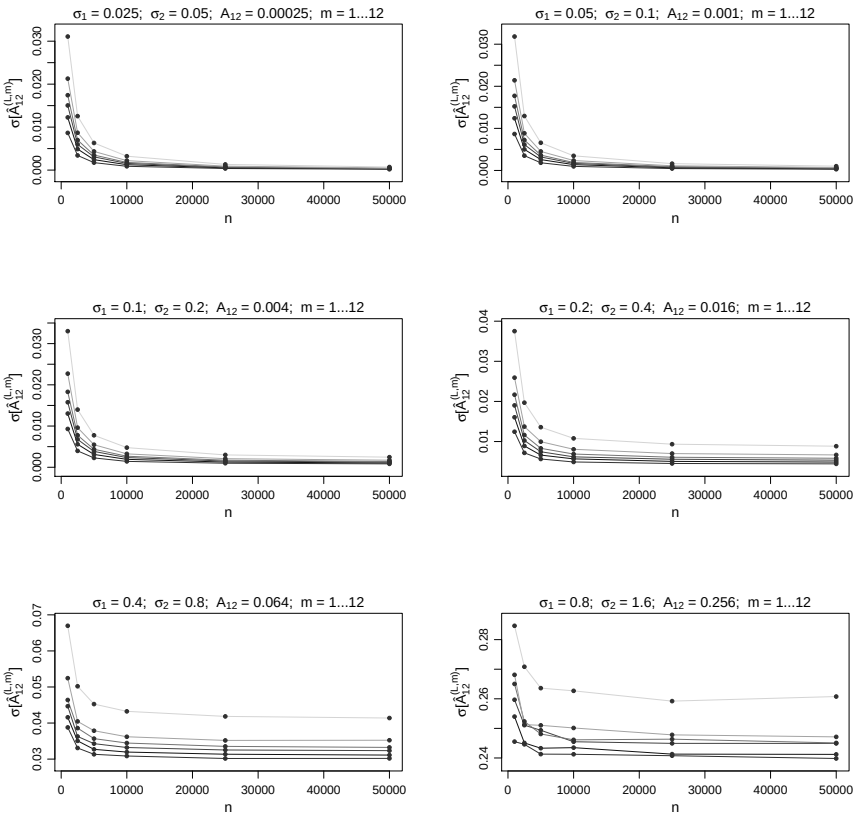

Figure 6: $\sigma\left[\hat{A}_{12}^{(L,m)}\right]$ as a function of n_h, considering decreasing m (from darker to brighter curve) and various choices of $\boldsymbol{\sigma}_\Gamma$.

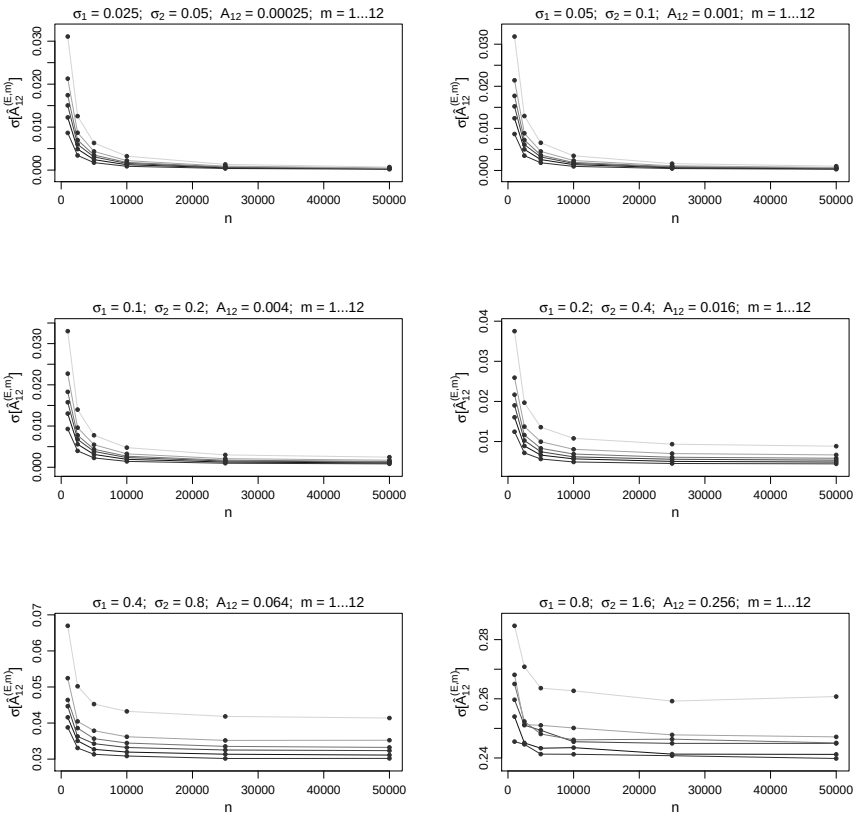

Figure 7: $\sigma\left[\hat{A}_{12}^{(E,m)}\right]$ as a function of n_h, considering decreasing m (from darker to brighter curve) and various choices of $\boldsymbol{\sigma}_\Gamma$.

6 An application to market data

We consider the hystorical series of bad loan rates from Bank of Italy. Using the quarterly hystorical series $TRI30529$ ($m = 4$) over a ten year period (from 1 Gen 2007 to 31 Dec 2017, $n = 10$, $\Delta t = 1$) we estimated $\hat{A}(0,1)$ applying equation (59) over a one-year period with respect to the different classes of total margin used (3 clusters). The result is reported in table 2. Since we know the hystorical number of risky subject $n_h(t)$ for each cluster $h = 1, \ldots, 3$ and observation date $t = \frac{1}{4}, \frac{2}{4}, \ldots, 10$, we are able to simulate the error $\sigma\left[\hat{A}^{(E,4)}(0,1)\right]$ and to compare it with $\sigma\left[\hat{A}^{(E,1)}(0,1)\right]$ showing that the latter is greater and so $\hat{A}^{(E,4)}(0,1)$ resulting to be a more precise estimator. In tables 3 and 4 errors are computed in the quasi-gaussian hypothesis. As expected, this is an understimate of the actual errors, since σ_k are not small enough to cope with the underlying hypothesis. The actual errors are estimated by numerical simulations and reported in tables 5 and 6. $\sigma\left[\hat{A}^{(E,4)}(0,1)\right]$ is confirmed to be a more precise observable than $\sigma\left[\hat{A}^{(E,1)}(0,1)\right]$.

0.01339845	0.01229858	0.01573644
0.01229858	0.01285280	0.01693868
0.01573644	0.01693868	0.02471498

Table 2: Estimator $\hat{A}^{(E,4)}(0,1)$ applied to the quarterly hystorical series $TRI30529$ over the period 1 Gen 2007 – 31 Dec 2017. The serie represent the bad loan rate over three different classes of total margin used.

0.003034152	0.002879915	0.003852332
0.002879915	0.002910586	0.003937250
0.003852332	0.003937250	0.005596842

Table 3: $\sigma\left[\hat{A}^{(E,4)}(0,1)\right]$ estimated with the quasi-gaussian approximation and the ansatz $n_h = \infty$.

0.02019638	0.02005440	0.02773821
0.02005440	0.02079141	0.02901729
0.02773821	0.02901729	0.04231579

Table 4: $\sigma\left[\hat{A}^{(E,1)}(0,1)\right]$ estimated with the quasi-gaussian approximation and the ansatz $n_h = \infty$.

The calibration of a CreditRisk$^+$ model can be accomplished by the eigenvalues decomposition of $\hat{A}^{(E,4)}(0,1)$, that in this case behaves as an exact SNMF, which is used in literature to solve this decomposition problem in general.

0.003951516	0.008113624	0.007855003
0.008113624	0.017565241	0.005271942
0.007855003	0.005271942	0.015905663

Table 5: $\sigma\left[\hat{A}^{(E,4)}(0,1)\right]$ estimated numerically, using the actual $n_h(t)$ observed for each quarter to take into account the binomial contribution to the error.

0.03004347	0.04092926	0.04966739
0.04092926	0.04292081	0.06119799
0.04966739	0.06119799	0.08094572

Table 6: $\sigma\left[\hat{A}^{(E,1)}(0,1)\right]$ estimated numerically, using the actual $n_h(t)$ observed for each year to take into account the binomial contribution to the error.

k	0	1	2	3
w_{1k}	0.0000000	0.2912034	0.29823104	0.4105656
w_{2k}	0.4814312	0.4486041	0.06996472	0.0000000
w_{3k}	0.5689608	0.2371924	0.00000000	0.1938468
σ_k		0.37517076	0.08324963	0.03475040

Table 7: The complete set of parameters $\{W, \boldsymbol{\sigma}_\Gamma\}$ necessary to specify the dependence structure in CreditRisk$^+$ model, obtained by the eigenvalues decomposition of $\hat{A}^{(E,4)}(0,1)$, as reported in Table 2.

References

[1] Credit Suisse First Boston, *CreditRisk$^+$, A Credit Risk Management Framework*,
www.csfb.com/institutional/research/assets/creditrisk.pdf

[2] Gundlach M., Lehrbass F. (eds.), *CreditRisk$^+$ in the Banking Industry*, Springer (2004).

[3] Klugman S.A., Panjer H.H., Willmot G.E., *Loss Models: From Data to Decisions*, Wiley (2012).

[4] Glasserman P., Li J., *Importance Sampling for Portfolio Credit Risk*, Management Science, 51 (11), Nov. 2005, pp. 1643-1656,
www0.gsb.columbia.edu/faculty/pglasserman/Other/is˙credit.pdf

[5] McNeil A., Frey R., Embrechts P., *Quantitative Risk Management*, Princeton University Press (2015).

[6] Mai J., Scherer M., *Simulating Copulas*, World Scientic, (2017).

[7] Vandendorpe A., Ho N., Vanduffel S., Van Dooren P., *On the parameterization of the CreditRisk$^+$ model for estimating credit portfolio risk*, Insurance: Mathematics and Economics, 42 (2008), pp. 736-745.