mathematics

MDPI

Article

Calibrating the CreditRisk$^+$ Model at Different Time Scales and in Presence of Temporal Autocorrelation †

Jacopo Giacomelli [1,2,*,†] and Luca Passalacqua [2,*]

[1] SACE S.p.A., Piazza Poli 42, 00187 Rome, Italy
[2] Department of Statistics, Sapienza University of Rome, Viale Regina Elena 295, 00161 Rome, Italy
* Correspondence: j.giacomelli@sace.it (J.G.); luca.passalacqua@uniroma1.it (L.P.)
† 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.

Abstract: The CreditRisk$^+$ model is one of the industry standards for the valuation of default risk in credit loans portfolios. The calibration of CreditRisk$^+$ requires, inter alia, the specification of the parameters describing the structure of dependence among default events. This work addresses the calibration of these parameters. In particular, we study the dependence of the calibration procedure on the sampling period of the default rate time series, that might be different from the time horizon onto which the model is used for forecasting, as it is often the case in real life applications. The case of autocorrelated time series and the role of the statistical error as a function of the time series period are also discussed. The findings of the proposed calibration technique are illustrated with the support of an application to real data.

Keywords: CreditRisk$^+$; calibration; time series; default correlation; dependence structure

MSC: 62F25; 62H12; 62H25; 62M10; 62P05

JEL Classification: C38; C51; G21; G22

Citation: Giacomelli, J.; Passalacqua, L. Calibrating the CreditRisk$^+$ Model at Different Time Scales and in Presence of Temporal Autocorrelation. *Mathematics* **2021**, *9*, 1679. https://doi.org/10.3390/math9141679

Academic Editor: Larissa Batrancea

Received: 4 June 2021
Accepted: 13 July 2021
Published: 16 July 2021

1. Introduction

While the development of modern portfolio credit risk models started in the 1980–1990 decade [1] within the framework of the Basel Accords, it is with the great credit crisis of 2008 [2] that increasing attention started to be paid to the precise determination of the structure of dependence among default events. It is well established [3] that tails of the distribution of the value of asset/liabilities portfolios are dominated by the structure of dependence rather than by the other fundamental components of credit risk (i.e., the marginal probability and the severity associated with each future default event). The vast research interest in modeling the structure of dependence resulted in the formalization of the so-called copula theory [4,5]. This "language" was explicitly adopted by the second generation of portfolio credit models to describe the dependence among loss events [6–9].

In this regard, the calibration issues raised by a particular structure of dependence (or, equivalently, the corresponding copula) can be as important as the choice of the structure itself. Generally, calibrating the dependence structure of a portfolio model is a demanding task, given the large number of parameters needed to provide a realistic description of the modeled dependencies, and considering that, on the other hand, historical 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: how to choose the frequency of the historical time series of default used to calibrate a classic credit portfolio model, CreditRisk$^+$, in order to provide the most accurate estimation of the structure of dependence parameters, or, in other words, how the calibration error "scales" with the time series

frequency. This problem is especially relevant for all the cases when the debtors underlying a credit portfolio are small/medium enterprises. The lack of market information, such as CDS spread, stock price, or bond yield, forces to calibrate the model using a reduced-form approach based on historical cluster data, such as default rate time series associated with the economic sector of each debtor. This case is typical in activities such as credit insurance, surety, and factoring. In most cases, publicly available time series have a sampling period ranging from one to three months (e.g., [10]), while the calibrated CreditRisk$^+$ model is used on a projection horizon that is at least one year long (e.g., the unwind period required to quantify a capital requirement both in Solvency 2 and in Basel 3 regulatory frameworks).

CreditRisk$^+$ [11], disclosed in 1997, belongs to the first generation of portfolio credit risk models of "actuarial inspiration". Applications of CreditRisk$^+$ to the credit insurance sector are documented in the literature well before the 2008 financial credit crisis [12,13], while research activity is still ongoing in the area of actuarial science [14]. At present, CreditRisk$^+$ is still one of the financial and actuarial industry standards for the assessment of credit risk in portfolios of financial loans or credit/suretyship policies.

Despite the vast research activity on this model and its calibration, the issue of using two different time scales for calibration and projection remains not investigated to date. The research conducted to date on the calibration of CreditRisk$^+$ [14] has addressed the issues related to the decomposition of a given covariance matrix among the time series, which is the final necessary step to complete the calibration of the model. However, the covariance matrix is obtained by the "classical" estimator, under the assumption that the sampling period of the time series and the projection horizon are equal.

This work shows that calibrating the model at a shorter time scale than the projection horizon is possible, nontrivial, and convenient. The internal consistency of the CreditRisk$^+$ assumptions when simultaneously imposed at different time scales has been proved and guarantees that the investigated calibration mode is not ill-posed. However, the form of the covariance estimator needed to obtain a set of parameters coherent with a specific projection horizon, using time series with a smaller sampling period, depends on the two chosen time scales. Indeed, the proposed estimator coincides with the classical one only when calibration and projection time scales are equal. Finally, we show that calibrating at a smaller time scale than the projection one provides a more precise estimation of the model parameters. The estimation error and its dependence on the difference between the two time scales are discussed.

The article is organized as follows. In Section 2, we summarize assumptions and features of the CreditRisk$^+$ model. In Section 3, we discuss the internal consistency of the model assumptions when imposing them to be simultaneously true at different time horizons. The calibration of the model parameters, which define the dependence structure, is considered in Section 4. The different degree of precision of the estimators defined at increasing time scales is discussed in Section 5. The techniques introduced in this work are applied to a real-world case study in Section 6. The main results are summarized in Section 7.

2. The CreditRisk$^+$ Model

The CreditRisk$^+$ model is a portfolio model developed by Credit Suisse First Boston (CSFB) by Tom Wilde [15] and coworkers, first documented in [11] and later widely discussed in [16]. 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, e.g., variation of the credit standing (the so-called "credit migration" effect). CreditRisk$^+$ can be classified as a frequency–severity model, cast in a single-period framework, with the peculiarity that a doubly-stochastic process (i.e., the Poisson–Gamma mixture) describes the frequency of default events. Loss severity is assumed to be deterministic, although this ansatz can be easily relaxed at the cost of some additional computational burden. However, severity-related issues can be neglected for what follows.

The structure of dependence of default events is described using a factor model framework, where factors are unobservable (i.e., latent) stochastic "market" variables, whose precise financial/actuarial identification is irrelevant since the model integrates on all possible realizations ("market scenarios"). Therefore, CreditRisk$^+$ can be further classified into the family of factor models and, in particular, into the subfamily 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 summarized 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 such that

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

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

$$L = \sum_{i=1}^{N} \mathbb{1}_i E_i \tag{2}$$

where each exposure E_i is supposed to be deterministic.

In order to ease the semianalytic 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.

Assumption 1 (CreditRisk$^+$ distributional assumption)**.** *Given a time horizon $(t, T]$ and a set of N risky debtors, the number Y_i of insolvency events generated by each i-th debtor over $(t, T]$ is distributed as follows:*

$$Y_i \sim \text{Poisson}(p_i(\mathbf{\Gamma})), \quad p_i(\mathbf{\Gamma}) \coloneqq q_i \cdot \left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik} \Gamma_k \right) \tag{3}$$

where $\mathbf{\Gamma} = (\Gamma_1 \ldots \Gamma_K) \in \mathbb{R}_+^K$ is an array of independent r.v.'s such that

$$\Gamma_k \sim \text{Gamma}\left(\beta_k^{-1}, \beta_k\right), \quad \beta_k \in \mathbb{R}_+ \tag{4}$$

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

$$\begin{aligned} &\omega_{ik} \geq 0, \qquad i = 1, \ldots, N, \quad k = 0, \ldots, K, \\ &\sum_{k=0}^{K} \omega_{ik} = 1, \quad i = 1, \ldots, N. \end{aligned} \tag{5}$$

The $\mathbf{\Gamma}$ parameters set $\{\beta_1 \ldots \beta_K\}$ is equivalent to the classical shape-scale parameterization $\{\alpha_k, \beta_k\}$ of each Gamma distributed r.v. Γ_k, after having imposed the assumption $\mathbf{E}[\Gamma_k] = 1$, that is stated in the original formulation of the CreditRisk$^+$ model. Hence, the k-th scale parameter β_k is equal to the variance σ_k^2 of Γ_k. Given the independence among Γ_k's, the covariance matrix Σ takes the form

$$\Sigma \coloneqq \mathbf{cov}[\mathbf{\Gamma}] = \mathbf{diag}\left(\sigma_1^2 \ldots \sigma_K^2\right) = \mathbf{diag}(\beta_1 \ldots \beta_K) \tag{6}$$

Assumption 1 implies that q_i is the unconditional expected default frequency

$$q_i = \mathbf{E}[p_i(\mathbf{\Gamma})] = \int_{\mathbb{R}_+^K} p_i(\mathbf{\Gamma})\, f(\mathbf{\Gamma}) d\Gamma_1 \ldots d\Gamma_K, \tag{7}$$

where

$$f(\boldsymbol{x}) = \prod_{k=1}^{K} \frac{x_k^{\alpha_k - 1}}{\beta_k^{\alpha_k} \Gamma(\alpha_k)} e^{-x_k/\beta_k}, \quad x_k \geq 0, \quad \alpha_k, \beta_k > 0, \tag{8}$$

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

$$\mathbf{E}[Y_i] = \mathbf{E}[\mathbb{1}_i] = q_i. \tag{9}$$

The portfolio loss is now represented by the r.v. L_Y

$$L_Y = \sum_{i=1}^{N} Y_i \cdot E_i, \quad \text{where} \quad Y_i | \boldsymbol{\Gamma} \sim \text{Poisson}(p_i(\boldsymbol{\Gamma})). \tag{10}$$

In [11], the distribution of L_Y is obtained by using a recursive method, further described in [17]. The accuracy, stability, and possible variants of the original algorithm are discussed in [16]. The same distribution can be easily computed through Monte Carlo simulation due to the availability of a dedicated importance sampling algorithm in [18].

Notice that, although the distributions of L and L_Y differ, the expected value of the portfolio loss is the same $\mathbf{E}[L] = \mathbf{E}[L_Y]$.

In the language of copula functions, the structure of dependence implied by (3) corresponds [19] 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., ([9] [§2.1])). The copula parameters are the factor loadings ω_{ik} and they can be gathered, taking into account the normalization condition stated in Assumption 1, in an $N \times K$ matrix Ω:

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

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

Remark 1. *This work is specifically focused on improving the estimation of the CreditRisk$^+$ copula parameters $\{\Omega, \Sigma\}$. Further investigations on the properties of CreditRisk$^+$ dependence structure, apart from those needed for the estimation improvement, and its comparison with the other copulae are beyond the scope of this study.*

As shown in [14], it holds

$$\mathbf{cov}[Y_i, Y_j] = q_i \, q_j \sum_{k=1}^{K} \omega_{ik} \omega_{jk} \sigma_k^2 + \delta_{ij} q_i, \tag{12}$$

where δ_{ij} is the Kronecker delta. Equation (12) 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, since the model is defined in a single-period framework, 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 (e.g., quarterly) in a consistent way, when calibrating the model parameters. Naively, it is reasonable to expect that the larger the information provided by the historical time series (i.e., the higher the frequency), the better the calibration. This issue is addressed in the next sections.

3. CreditRisk$^+$ Using Multiple Unwind Periods

The original CreditRisk$^+$ formulation, summarized in Assumption 1, defines the model in a uniperiodal framework, where only one time scale $T - t$ is considered. In this section, we discuss the internal consistency of the model assumption when imposing it more than once at distinct time scales. In this context, the expression "internal consistency" means that it is possible and well-posed imposing Assumption 1 to be true at two distinct time scales. The same applies also considering a slightly modified version of the CreditRisk$^+$ framework (i.e., imposing Assumption 2, introduced in the following, instead of Assumption 1).

Extending the original CreditRisk$^+$ formulation to a multiperiod framework enables the calibration of the model considering a time scale different from the one chosen for its application. The results presented in this section are applied in the next Section 4 to estimate the elements of the matrix

$$A \coloneqq \Omega^T \Sigma \Omega. \tag{13}$$

Estimating A is a fundamental step in order to complete the calibration of the model. In Section 4 estimators are defined using historical series sampled with a period that is not necessarily equal to the projection horizon on which Σ and Ω are defined. Section 5 shows the convenience of choosing a sampling period shorter than the projection horizon in order to evaluate $\hat{A}$.

3.1. The Single Unwind Period Case

As discussed in Section 2, in CreditRisk$^+$ each risk (i.e., debtor) is modeled by a Poisson distributed r.v. Y_i, although the Bernoulli distribution is the natural choice to represent absorbing events, such as default. Assumption 1 is convenient in terms of analytical tractability since L_Y distribution can be computed through a semianalytical method. However, in order to address the problem of calibrating CreditRisk$^+$ in a "roll-over" framework, defined by an arbitrary set of time intervals, it is useful to recover the Bernoulli representation of each debtor by introducing a new r.v. $\widetilde{Y}_i \coloneqq \mathbb{1}_{Y_i>0}$.

Both the r.v. Y_i and its distribution parameter $p_i(\mathbf{\Gamma})$ can take values larger than 1. This is formally correct, given that $Y_i \sim \text{Poisson}(p_i(\mathbf{\Gamma}))$, despite not coping with the representation of absorbing events, that can occur at most once by definition. The so-called "Poisson approximation", introduced by substituting $\mathbb{1}_i$ with Y_i, is numerically sound as q_i approaches to zero—a condition that is well fulfilled in most real world relevant cases.

Indeed, Assumption 1 implies that $\widetilde{Y}_i|\mathbf{\Gamma} \sim \text{Bernoulli}(\tilde{p}_i(\mathbf{\Gamma}))$ where the distribution parameter is

$$\tilde{p}_i(\mathbf{\Gamma}) = \text{Prob}(Y_i > 0|\mathbf{\Gamma}) = 1 - \exp\left[-q_i\left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik}\,\Gamma_k\right)\right]. \tag{14}$$

It holds by construction

$$\mathbf{E}\left[\widetilde{Y}_i\right] = \int_{\mathbb{R}_+^K} \tilde{p}_i(\mathbf{\Gamma}) f(\mathbf{\Gamma}) d\Gamma_1 \dots d\Gamma_K. \tag{15}$$

Computing the integral in (15) and then approximating the term $\exp[-q_i\omega_0]$ with its second order Taylor series centered at $q_i = 0$ leads to the following result.

Proposition 1 (Asymptotic equivalence between Bernoulli and Poisson representation of risks)**.** *Let $\widetilde{Y}_i \coloneqq \mathbb{1}_{Y_i>0}$ where Y_i is distributed according to Assumption 1. Then*

$$\tilde{q}_i \coloneqq \mathbf{E}\left[\widetilde{Y}_i\right] = 1 - e^{-q_i\,\omega_{i0}} \prod_{k=1}^{K}\left(1 + q_i\,\omega_{ik}\sigma_k^2\right)^{-1/\sigma_k^2}. \tag{16}$$

Further,

$$\tilde{q}_i = q_i + \mathcal{O}(q_i^2) \xrightarrow{q_i \to 0^+} q_i. \tag{17}$$

Proposition 1 implies that $\mathbf{E}[L_{Y_i}] \simeq \mathbf{E}[L_{\widetilde{Y}_i}]$, provided that $q_i \ll 1$. Moreover, the same result enables also the exact satisfaction of $\mathbf{E}\left[L_{Y_i}\right] = \mathbf{E}\left[L_{\widetilde{Y}_i}\right]$, in case the stochastic parameter $\tilde{p}_i(\mathbf{\Gamma})$ is redefined through the substitution $q_i \mapsto q_i'$, where q_i' verifies the following modified version of (16):

$$\mathbf{E}\left[\widetilde{Y}_i(q_i')\right] = 1 - e^{-q_i' \omega_{i0}} \prod_{k=1}^{K} \left(1 + q_i' \, \omega_{ik} \sigma_k^2\right)^{-1/\sigma_k^2} = q_i = \mathbf{E}[Y_i]. \tag{18}$$

It is worth noticing that the substitution $\mathbb{1}_i \mapsto Y_i$ discussed in Section 2 implies the preservation of the expected value $\mathbf{E}[L] = \mathbf{E}[L_Y]$ due to the fact that it is done before the introduction of the market factors $\mathbf{\Gamma}$. On the other hand, restoring the Bernoulli representation of each risk after having introduced the dependence structure requires the results presented in Proposition 1.

Proposition 1 permits the introduction of a slightly modified version of the CreditRisk$^+$ model that is asymptotically equivalent to the original one stated in Assumption 1. The equivalence between the two models is further analyzed in the next sections.

Assumption 2 (Modified CreditRisk$^+$ distributional assumption)**.** *Given a time horizon $(t, T]$ and a set of N risky debtors, the number of insolvency events generated by each i-th debtor over $(t, T]$ is represented by the r.v. $\widetilde{Y}_i \sim$ Bernoulli$(\tilde{p}_i(\mathbf{\Gamma}))$, where the distribution parameter $\tilde{p}_i(\mathbf{\Gamma})$ satisfies (14). Assumptions on market factors $\mathbf{\Gamma}$ and factor loadings Ω remain the same stated in Assumption 1.*

In Assumption 2 the linear dependence of the parameters $p_i(\mathbf{\Gamma})$ from the latent variables has been replaced with a log link function. Thus, the modified version of CreditRisk$^+$ is also referred to as "exponential" in the following.

3.2. The Multiple Unwind Periods Case

This section investigates the consequences of imposing the internal consistency of Assumption 1 or Assumption 2 at distinct time scales. Assumptions 3 and 4 are introduced hereinafter, in order to specify the family of parameters that have to be considered at the distinct time intervals where the model is applied.

The following assumption guarantees the internal consistency at different time scales of the classical CreditRisk$^+$ model, defined in Assumption 1.

Assumption 3 (CreditRisk$^+$ parameters at different time scales)**.** *Let $t \equiv t_0, t_1, \ldots, t_m \equiv T$ be a partition of the time interval $(t, T]$. Let Assumption 1 be satisfied over each j-th interval $(t_{j-1}, t_j]$ by the set $\{Y_i^{(j)}\}$ $(i = 1 \ldots N)$, where $Y_i^{(j)}$ is the r.v. representing the i-th risk observed during the j-th interval and the following holds for the associated set $\{q_i^{(j)}; \mathbf{\Gamma}^{(j)}; \Omega^{(j)}\}$ of parameters and market factors:*

$$q_i^{(j)} = q_i \frac{t_j - t_{j-1}}{T - t} = \text{constant}, \tag{19}$$

$$\Gamma_k^{(j)} \sim \text{Gamma}\left(\sigma_k^{-2} \, \xi_{kj}^{-1} \, \tfrac{t_j - t_{j-1}}{T-t}, \sigma_k^2 \, \xi_{kj} \, \tfrac{T-t}{t_j - t_{j-1}}\right), \tag{20}$$

$$\Omega^{(j)} = \Omega, \tag{21}$$

where $\xi_{kj} \in \mathbb{R}_+$.

Further, the following assumption guarantees the internal consistency at different time scales of the modified version of CreditRisk$^+$ model, introduced in Assumption 2.

Assumption 4 (Modified CreditRisk$^+$ parameters at different time scales)**.** *Let $t \equiv t_0, t_1, \ldots, t_m \equiv T$ be a partition of the time interval $(t, T]$. Let Assumption 2 be satisfied over each j-th interval $(t_{j-1}, t_j]$ by the set $\{\widetilde{Y}_i^{(j)}\}$ $(i = 1 \ldots N)$, where $\widetilde{Y}_i^{(j)}$ is the r.v. representing the i-th risk observed during the j-th interval. The associated set $\{q_i^{(j)}; \mathbf{\Gamma}^{(j)}; \Omega^{(j)}\}$ of parameters and market factors satisfies the same assumptions stated in Assumption 3.*

Finally, for the sake of simplicity, the additional Assumption 5 is introduced, with regard to the independence among market factors considered at different times. However, being possible that real-data time series violate Assumption 5, this assumption is weakened in the following Section 3.3.

Assumption 5 (Non-autocorrelated market factors)**.** *Given Assumption 3, let*

$$\mathbf{cov}\left[\Gamma_k^{(j)}, \Gamma_k^{(j')}\right] = \delta_{jj'}\mathbf{var}\left[\Gamma_k^{(j)}\right]. \tag{22}$$

Considering the assumptions introduced above, we prove that CreditRisk$^+$ is internally consistent when extended to a roll-over framework.

Theorem 1 (Internal consistency of CreditRisk$^+$ in absence of autocorrelation)**.** *Let us consider a set of risks $\{Y_i\}$ $(i = 1 \ldots N)$, observed through a time horizon $(t, T]$, and an arbitrary partition $t \equiv t_0, t_1, \ldots, t_m \equiv T$ of $(t, T]$, such that Assumptions 3 ("CreditRisk$^+$ parameters at different time scales") and Assumption 5 ("non-autocorrelated market factors") are verified with*

$$\xi_{kj} = 1 \tag{23}$$

for each $i = 1 \ldots N$, $k = 1 \ldots K$ and $j = 1 \ldots m$. Then $\{Y_i\}$ satisfies Assumption 1 ("CreditRisk$^+$ distributional assumption") over $(t, T]$.

The statement above remains true replacing Assumption 3 with Assumption 4 ("modified CreditRisk$^+$ parameters at different time scales") and Assumption 1 with Assumption 2 ("modified CreditRisk$^+$ distributional assumption"), ceteris paribus.

The proof of Theorem 1 is reported in Appendix A.1.

This result shows that extending the CreditRisk$^+$ model to a multiperiod framework is well-posed.

Remark 2. *The choice $\xi_{jk} = 1$ implies no loss of generality, since a different (positive) constant $\xi_{jk} = c$ is equivalent to redefine the variances of the market factors $c\sigma_k^2 \mapsto \sigma_k^2$.*

3.3. Internal Consistency and Autocorrelation in Time Series

As shown in Section 2, the dynamics of each parameter p_i is induced by the latent Gamma factors only. Imposing Assumption 5 to any (arbitrarily short) time scale implies that considered time series $\{\Gamma_k^{(j)}\}_{j=1,2,\ldots}$ must exhibit zero autocorrelation. Hence autocorrelation must be completely absent from the historical default frequencies too.

However, this requirement could not be satisfied by the observed time series used in calibrating the model. Indeed, we need to verify that the model can preserve its internal consistency if autocorrelation has to be considered.

The purpose of this work is to investigate whether it is possible and convenient to calibrate the CreditRisk$^+$ model at a time scale that copes with the available historical data (i.e., the sampling period of the historical time series) instead of using the same time scale needed for projections (usually bigger). Hence, in case it is not possible to preserve the internal consistency of the model at each arbitrary time scale, due to the presence of autocorrelation, it is sufficient to ask that it holds up to the smallest of the two time scales of interest—the historical sampling period and the projection horizon.

Let us specialize to the constant mesh case $t_j - t_{j-1} = (T-t)/m = \delta_m$. This choice copes with a typical real case, where the sampling period δ_m of the available historical time series is constant and the considered projection horizon $T - t$ is a multiple of it. Under these premises, a weakened version of Assumption 5 is introduced.

Assumption 6 (Autocorrelated market factors)**.** *Given Assumption 3, for each k-th latent variable, considered at the time scale δ_m, a time-invariant ACF ϱ_{xk} exists, such that*

$$\mathbf{cov}\left(\Gamma_k^{(j)}, \Gamma_k^{(j+x)}\right) = \varrho_{xk}\mathbf{var}\left(\Gamma_k^{(j)}\right). \tag{24}$$

Furthermore, the following closure with respect to the addition holds

$$\sum_{j=1}^{m} \Gamma_k^{(j)} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{25}$$

for a couple α_k, β_k of shape and scale parameters.

Assumption 6 is considered instead of Assumption 5 to state the following alternate version of Theorem 1.

Theorem 2 (Internal consistency of CreditRisk$^+$ model in presence of autocorrelation)**.** *Let us consider a set of risks $\{Y_i\}$ $(i = 1 \dots N)$, observed through a time horizon $(t, T]$, and a uniform partition $\{t_j := t + j\delta_m\}_{j=1\dots m}$ of $(t, T]$, such that Assumption 3 ("CreditRisk$^+$ parameters at different time scales") and Assumption 6 ("autocorrelated market factors") are verified with*

$$\xi_{kj} = \left[1 + 2\sum_{x=1}^{m-1} \varrho_{xk}\left(1 - \frac{x}{m}\right)\right]^{-\frac{1}{2}} \tag{26}$$

for each $i = 1 \dots N$, $k = 1 \dots K$ and $j = 1 \dots m$. Then $\{Y_i\}$ satisfies Assumption 1 ("CreditRisk$^+$ distributional assumption") over $(t, T]$.

The statement above remains true replacing Assumption 3 with Assumption 4 ("modified CreditRisk$^+$ parameters at different time scales") and Assumption 1 with Assumption 2 ("modified CreditRisk$^+$ distributional assumption"), ceteris paribus.

The proof of Theorem 2 is reported in Appendix A.2.

Assumption 6 can be either well-posed or ill-posed, depending on the considered ϱ_{xk}. The trivial case $\varrho_{xk} = 0$ for each $x \in \mathbb{Z}$ copes with Assumption 5. Correlated Gamma variables, as well as the distributional properties of the sum of Gamma variables, have been intensively studied in the literature, and this is still an active research field [20–23], due to its relevance for information technology. At least in case of identically distributed Gamma variables—such as $\Gamma_k^{(j)}$ in our framework—with ACF obeying to a power-law

$$\varrho_{xk} = \rho_k^{|x|}, \quad \rho_k \in (0, 1), \tag{27}$$

the distribution of the sum Γ_k is known to be approximately Gamma [20], while more generical cases imply the sum to be distributed differently [22,23]. Moreover, it is known that partial sums of independent Gamma variables can be used to generate sequences of (auto)correlated Gamma variables [21].

Remark 3. *The exponential ACF in Equation (27) provides a non-trivial case that satisfies Assumption Assumption 6 and, thus, Theorem 2. In the following Section 4.4, Theorem 2 permits the estimation of A in presence of autocorrelated time series. Equation (27) is then considered in Section 5.3 to investigate numerically the estimators introduced in Section 4.4. However, to date,*

a general framework is missing to tell whether a given ϱ_{xk} lets the partial sums $\sum_j \Gamma_k^{(j)}$ remain (approximately) Gamma distributed, with the exception of exponential ACFs.

The estimators introduced in Section 4.4 to consider autocorrelation in time series are still applicable to an inconsistent framework, provided that at least the latent variables Γ_k (defined onto the projection horizon) are Gamma distributed and $\Gamma_k^{(j)}$ satisfy the mean and variance requirements implied by Assumption 6 above.

4. Calibration of the Structure of Dependence

The model is calibrated based on a partition of the risks in H homogeneous sets $c_h(t), h = 1, \ldots, H$. In this context "homogeneity" means that two 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 occurrence of defaults. On the contrary, the structure of dependence, defined by $\boldsymbol{\omega}^{(h)}$ is supposed to be time-independent.

Hence, solving the calibration problem requires the evaluation of

- H factor loading vectors $\{\boldsymbol{\omega}^{(h)}\}_{h=1\ldots H}$, that link each of the homogenous clusters to the K latent variables;
- K volatilities $\{\sigma_k\}_{k=1\ldots K}$, needed to specify the distribution of each of the latent variables.

The calibration is achievable by a two-step procedure. Firstly, the matrix $A := \Omega^T \Sigma \Omega$, introduced in Section 3, is estimated. Then, A is decomposed under the proper constraints in order to evaluate Ω and Σ separately. This section describes a method to complete the first step, providing an estimator of A both for the single and the multiple unwind period cases, with a moment-matching approach that allows expressing $\hat{A}$ as a function of the covariance matrix among the historical frequencies of default. The second step is addressed later in Section 6, which provides an example of calibration using a real data set.

Adopting the standard CreditRisk$^+$ Assumption 1, Equation (12) can be used to link the covariance matrix among the historical frequencies of default with the matrix A. In Section 4.1, $\hat{A}$ is provided in the case of historical frequencies of default, sampled with the same tenor of the projection horizon. In Section 4.2, $\hat{A}$ is generalized to the case of historical frequencies of default sampled with an arbitrary tenor.

Furthermore, in Section 4.3, $\hat{A}$ is determined under the exponential version of the CreditRisk$^+$ framework, introduced in Assumption 2. Thanks to this modified assumption, the corresponding functional form of $\hat{A}$ is simpler than the one obtained in Section 4.2 based on Assumption 1.

Sections 4.2 and 4.3 cope with Assumption 5, that implies absence of autocorrelation in time series. The final Section 4.4 uses Assumption 6 instead, generalizing the main results presented in this section to the case where autocorrelation must be taken into account. In this case, the simpler form of $\hat{A}$ obtained in Section 4.3 comes in handy in the generalization to the non-trivial ACF case.

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 $n_h(t) := |c_h(t)|$, $F_h := \frac{1}{n_h(t)} \sum_{i \in c_h(t)} Y_i$ and $G_h := 1 - F_h$. The expected values of F_h and G_h are respectively:

$$q_h := \mathbf{E}[F_h] = \frac{\sum_{i \in c_h(t)} q_i}{n_h(t)}, \tag{28}$$

$$s_h := \mathbf{E}[G_h] = 1 - \mathbf{E}[F_h]. \tag{29}$$

Remark 4. *The slight abuse of notation in (28) is done to avoid the introduction of a new symbol to represent $\mathbf{E}[F_h]$. However, the letters chosen for indexing risks and cluster ("i" and "h" respectively) are maintained in the following of this work, clarifying the meaning of the "q" symbol each time it is used.*

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

$$\begin{aligned}\mathbf{cov}(F_h, F_{h'}) &= \mathbf{E}[(F_h - \mathbf{E}[F_{h'}])(F_{h'} - \mathbf{E}[F_{h'}])] \\ &= \frac{1}{n_h n_{h'}} \mathbf{E}\left[\sum_{i \in c_h} (Y_i - q_i) \sum_{i' \in c_{h'}} (Y_{i'} - q_{i'})\right] \\ &= \frac{1}{n_h n_{h'}} \sum_{i \in c_h} \sum_{i' \in c_{h'}} \mathbf{cov}(Y_i, Y_{i'}),\end{aligned} \tag{30}$$

that, using Equation (12), becomes:

$$\mathbf{cov}(F_h, F_{h'}) = \frac{1}{n_h n_{h'}} \sum_{i \in c_h} \sum_{i' \in c_{h'}} \left(q_i q_{i'} \sum_{k=1}^{K} \omega_{ik} \omega_{i'k} \sigma_k^2 + \delta_{ii'} q_i \right). \tag{31}$$

Equation (31) shows the relation between the observed covariance of default frequencies and the factor loadings, describing the structure of dependence of the model.

Moreover, assuming that all risks in a given homogenous set share the same factor loadings, the above expression simplifies to:

$$\mathbf{cov}(F_h, F_{h'}) = q_h q_{h'} \sum_{k=1}^{K} \omega_{hk} \omega_{h'k} \sigma_k^2 + \delta_{hh'} \frac{q_h}{n_h} \tag{32}$$

Notice that the second term in Equation (32) is present only when $h = h'$, and becomes quickly negligible as n_h grows (since $q_h < 1$).

Equation (32) enables the estimation of A over the same time scale $T - t$ used for projections:

$$\hat{A}_{hh'} = \frac{1}{q_h q_{h'}} \left[\hat{\mathbf{cov}}(F_h, F_{h'}) - \delta_{hh'} \frac{q_h}{n_h} \right]. \tag{33}$$

4.2. The Multiple Unwind Period Case

Let us consider a set of H time series defined using a constant step $\delta_m = (T - t)/m$. As done in Section 2, each variable introduced in Section 4.1 for the time interval $(t, T]$ can be redefined over each of the considered time intervals. Namely, in the following we use the set of observables quantities $\{F_h, G_h, q_h, s_h\}$, measured either over $(t, T]$ or $(t_{j-1}, t_j = t_{j-1} + \delta_m]$ or a generic time interval $(t, t']$. For the latter two cases, we introduce the notation $\{F_h^{(j)}, G_h^{(j)}, q_h^{(j)}, s_h^{(j)}\}$ and $\{F_h(t, t'), G_h(t, t'), \dots\}$, respectively. Further, the variables

$$F_{mh} := 1 - \prod_{j=1}^{m} \left[1 - F_h^{(j)}\right], \tag{34}$$

$$G_{mh} := \prod_{j=1}^{m} G_h^{(j)} = 1 - F_{mh} \tag{35}$$

are introduced.

In CreditRisk$^+$, $F_h(t, t')$ arises from a doubly stochastic process, since each absorbing event is generated conditioned to the latent systematic factors. For the sake of simplicity, we neglect the idiosyncratic uncertainty brought by each $Y_i(t, t')$. In fact, for $n_h(t)$ large enough, the Bernoulli (or Poisson) r.v.'s contributions to the variance of $F_h(t, t')$ are dominated by the contribution of $\mathbf{\Gamma}(t, t')$. This permits the following assumption.

Assumption 7 (Large clusters)**.** *For each cluster c_h ($h = 1 \dots H$) and each time interval $(t, t'] \subseteq (t, T]$ it holds*

$$\mathbf{var}\left[F_h(t, t') | \mathbf{\Gamma}(t, t')\right] = 0\,.$$

Then the following holds:

Proposition 2 (CreditRisk$^+$ scale-invariance law)**.** *Let us consider a set of risks* $\{Y_i\}$ $(i = 1 \dots N)$*, observed through a time horizon* $(t_a, t_b]$ *and classified into a set of homogenous clusters* c_h $(h = 1 \dots H)$*. Let Assumption 3 ("CreditRisk$^+$ parameters at different time scales"), Assumption 5 ("non-autocorrelated market factors") and Assumption 7 ("large clusters") hold with* $\xi_{kj} = 1$ *for each* $(t, T] \subseteq (t_a, t_b]$ *and for each uniform partition* $t \equiv t_0 < t_1 < \cdots < t_m \equiv T$ *of* $(t, T]$*,* $(m \in \mathbb{N}^*)$*. Then, the couple* $F_h(t, T), F_{h'}(t, T)$ *satisfies the conservation law*

$$[\mathbf{cov}(F_h(t,T), F_{h'}(t,T)) + s_h(t,T)s_{h'}(t,T)]^{\frac{1}{T-t}} = \text{constant}. \tag{36}$$

for each pair of clusters $c_h, c_{h'}$ *and each* $(t, T] \subseteq (t_a, t_b]$.

The proof of Proposition 2 is reported in Appendix A.3.

Proposition 2 is one of the main results of this work. It allows to build an estimator of $\mathbf{cov}(F_h(t,T), F_{h'}(t,T))$ using default frequencies $F_h^{(j)}$ defined on a different time scale δ_m. The dependence upon m of the precision of the covariance estimator is discussed in Section 5.

Indeed, applying Proposition 2 to Equation (33), it is possible to calibrate the dependence structure of the CreditRisk$^+$ model, by first determining the elements of the A matrix as

$$A_{hh'} = \frac{1}{q_h q_{h'}} \left[\left(\mathbf{cov}\left(F_h^{(j)}, F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)} \right)^m - s_h s_{h'} - \delta_{hh'} \frac{q_h}{n_h} \right] \tag{37}$$

for any $j = 1, \dots, m$, and then decomposing A, thus obtaining a separate estimate of the $\{\Omega, \sigma_\Gamma^2\}$ parameters. The SNMF decomposition can be performed, e.g., by using the technique described in [14].

4.3. The Exponential Case

In this section the problem of calibrating the dependence structure is addressed using the exponential form of the model introduced in Assumptions 2 and 4. Theorem 1 proves that also the exponential form remains consistent when considering multiple unwind periods. Since now $\widetilde{Y}_i$ variables are used instead of the corresponding Y_i, the frequencies F_h and their complements G_h are replaced by $\widetilde{F}_h$ and $\widetilde{G}_h$, defined by the substitution $Y_i \mapsto \widetilde{Y}_i$ in F_h and G_h definitions, respectively. Furthermore, it is convenient to introduce the following

$$L_h := -\frac{q_h}{q_h^\star} \ln \widetilde{G}_h \tag{38}$$

where

$$q_h^\star := -\ln \frac{\sum_{i \in c_h(t)} e^{-q_i}}{n_h(t)}. \tag{39}$$

The notation introduced in Section 4.2 for $\{F_h, G_h, q_h, \dots\}$ is extended to the exponential case as well. Hence, the sets of symbols $\{\widetilde{F}_h(t,t'), \widetilde{G}_h(t,t'), \dots\}$ and $\{\widetilde{F}_h^{(j)}, \widetilde{G}_h^{(j)}, \dots\}$ are also used. The log link function that relates $\tilde{p}_i$ and $\mathbf{\Gamma}$ simplifies the form of the scale invariariance law presented in Proposition 2. Indeed, in this case the following holds.

Proposition 3 (Modified CreditRisk$^+$ scale-invariance law)**.** *Let us consider a set of risks* $\{\widetilde{Y}_i\}$ $(i = 1 \dots N)$*, observed through a time horizon* $(t_a, t_b]$ *and classified into a set of homogenous clusters* c_h $(h = 1 \dots H)$*. Let Assumption 4 ("modified CreditRisk$^+$ parameters at different time scales"), Assumption 5 ("non-autocorrelated market factors") and Assumption 7 ("large clusters") hold with* $\xi_{kj} = 1$ *for each* $(t, T] \subseteq (t_a, t_b]$ *and for each uniform partition* $t \equiv t_0 < t_1 < \cdots < t_m \equiv T$ *of* $(t, T]$*,* $(m \in \mathbb{N}^*)$*. Then* $L_h(t,T), L_{h'}(t,T)$ *obey to the conservation law*

$$\frac{1}{T-t} \mathbf{cov}[L_h(t,T), L_{h'}(t,T)] = \text{constant} \tag{40}$$

for each pair of clusters $c_h, c_{h'}$ *and each* $(t, T] \subseteq (t_a, t_b]$.

The proof of Proposition 3 is reported in Appendix A.4.

Proposition 3 states a conservation law for the modified version of the model, likewise Proposition 2 in the original (i.e., Poisson–Gamma) CreditRisk$^+$ framework. The form obtained for the LHS of Equation (40) is simpler than the corrisponding LHS of Equation (36). In general, this framework results to be more tractable than the original model. This is especially useful when estimating A given a non-trivial ACF, as shown in the next Section 4.4.

In this case, A can be estimated as

$$A_{hh'} = \frac{1}{q_h q_{h'}} \mathbf{cov}[L_h, L_{h'}] = \frac{1}{q_h^\star q_{h'}^\star} \mathbf{cov}\left[\ln\left(1 - \widetilde{F}_h\right), \ln\left(1 - \widetilde{F}_{h'}\right)\right] \tag{41}$$

where we have neglected the contribution of $\mathbf{cov}(\widetilde{Y}_i, \widetilde{Y}_i) \propto \frac{1}{n_h(t_1)} \simeq 0$. Definition (38) and Proposition 3 imply

$$A_{hh'} = \frac{m}{q_h^{\star(j)} q_{h'}^{\star(j)}} \mathbf{cov}\left[\ln\left(1 - \widetilde{F}_h^{(j)}\right), \ln\left(1 - \widetilde{F}_{h'}^{(j)}\right)\right] \tag{42}$$

for each $j = 1 \dots m$.

4.4. Handling Autocorrelated Time Series in Calibration

In this section a generalization of estimators in Equations (37) and (42) is provided, in case Assumption 5 has to be replaced with Assumption 6 due to the presence of autocorrelation in time series. We preliminarily report below a second order approximation that comes in handy to generalize Equation (37).

$$\begin{aligned} \prod_{j=1}^{m} \mathbf{E}\left[G_h^{(j)} G_{h'}^{(j)}\right] &= \prod_{j=1}^{m} \left(1 - q_h^{(j)} - q_{h'}^{(j)} + \mathbf{E}\left[F_h^{(j)} F_{h'}^{(j)}\right]\right) \\ &= 1 - \sum_{j=1}^{m} \left(q_h^{(j)} + q_{h'}^{(j)}\right) + \sum_{j=1}^{m} \mathbf{E}\left[F_h^{(j)} F_{h'}^{(j)}\right] \\ &+ \sum_{j<j'} \sum_{h,h'=1,2} q_h^{(j)} q_{h'}^{(j')} + \dots \end{aligned} \tag{43}$$

We now consider again the relation between $\mathbf{cov}(F_{mh}, F_{mh'})$ and $\mathbf{cov}\left(F_h^{(j)} F_{h'}^{(j)}\right)$ implied by Proposition 2, under the presence of autocorrelation for the latent variables. Unlike in Section 3.2, in this case covariance terms at delay $|j - j'| \geq 1$ cannot be nullified.

$$\begin{aligned} \mathbf{cov}(F_{mh}, F_{mh'}) &= \mathbf{E}\left[\prod_{j=1}^{m} G_h^{(j)} G_{h'}^{(j)}\right] - s_h s_{h'} \\ &= 1 - \sum_{j=1}^{m} \left(q_h^{(j)} + q_{h'}^{(j)}\right) \\ &+ \sum_{j<j'} \sum_{h,h'=1,2} \mathbf{E}\left[F_h^{(j)} F_{h'}^{(j')}\right] \\ &+ \sum_{j=1}^{m} \mathbf{E}\left[F_h^{(j)} F_{h'}^{(j)}\right] - s_h s_{h'} + \dots \end{aligned} \tag{44}$$

Replacing Equation (43) into Equation (44), we have

$$\begin{aligned}\mathbf{cov}(F_{mh}, F_{mh'}) &= \prod_{j=1}^{m} \mathbf{E}\left[G_h^{(j)} G_{h'}^{(j)}\right] \\ &+ \sum_{j<j'} \sum_{h,h'=1,2} \mathbf{cov}\left[F_h^{(j)} F_{h'}^{(j')}\right] - s_h s_{h'} + O_3 \end{aligned} \tag{45}$$

where O_3 is a compact notation for the sum of all the terms of order 3 or greater. Given that $O_3 \xrightarrow{\mathbf{q}\to\mathbf{0}} 0$, the approximation $O_3 \approx 0$ is numerically sound in practice and implies the following generalization of $A_{hh'}$ in Equation (37)

$$A_{hh'} \approx \frac{1}{q_h q_{h'}} \left[\left(\mathbf{cov}\left(F_h^{(j)}, F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)} \right)^m + AC_{hh'}^{(L)} - s_h s_{h'} - \delta_{hh'} \frac{q_h}{n_h} \right], \tag{46}$$

where the autocorrelation term $AC^{(L)}$ is defined as

$$AC_{hh'}^{(L)} := \sum_{x=1}^{m-1} (m-x) \left(\mathbf{cov}\left[F_h^{(j)} F_h^{(j+x)}\right] + \mathbf{cov}\left[F_{h'}^{(j)} F_{h'}^{(j+x)}\right] + 2\mathbf{cov}\left[F_h^{(j)} F_{h'}^{(j+x)}\right] \right). \tag{47}$$

This completes the extension of the linear case presented in Section 4.2 to autocorrelated time series.

The exponential case—introduced in Section 4.3—turns out to be more tractable, since the linear structure implied by Proposition 3 allows us to avoid approximations similar to the one applied to extend the linear case above. Indeed, only the simplification implied by Assumption 5 must be abandoned, implying

$$\mathbf{cov}[L_h, L_{h'}] = m\, \mathbf{cov}\left[L_h^{(j)}, L_{h'}^{(j)}\right] + \sum_{x=1}^{m-1} 2(m-x) \mathbf{cov}\left[L_h^{(j)}, L_{h'}^{(j+x)}\right]. \tag{48}$$

This is implied by the fact that $L_h^{(j)}$ are still identically distributed for the same h but not independent. Hence, the estimator in Equation (42) becomes

$$A_{hh'}(t,T) = \frac{m}{q_h^{\star(j)} q_{h'}^{\star(j)}} \mathbf{cov}\left[\ln\left(1-\widetilde{F}_h^{(j)}\right), \ln\left(1-\widetilde{F}_{h'}^{(j)}\right)\right] + AC_{hh'}^{(E)} \tag{49}$$

where

$$AC_{hh'}^{(E)} := \frac{1}{q_h^{\star(j)} q_{h'}^{\star(j)}} \sum_{x=1}^{m-1} 2(m-x) \mathbf{cov}\left[\ln\left(1-\widetilde{F}_h^{(j)}\right), \ln\left(1-\widetilde{F}_{h'}^{(j+x)}\right)\right] \tag{50}$$

5. The Advantage of a Short Sampling Period

Let us consider a Δ_t-long projection period 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 $5 \leq n \leq 20$. Moreover, let the historical time series be sampled with a period δ_m, which is m times smaller than Δ_t (i.e., $\delta_m := \Delta_t / m$). Considering $\Delta_t = 1$ year, realistic assumptions are $m = 4$ (quarterly time series) or $m = 12$ (monthly time series). Therefore, the considered time series are defined over $m \times n$ intervals of length δ_m, defined by a schedule $t_0, \ldots, t_{m\times n}$.

This section discusses the precision improvement achievable by calibrating the model on historical default time series with a period smaller than the time horizon on which the calibrated model is applied. Indeed, the statistical error on the determination of A depends on m, i.e., on the sampling frequency of the observations, as shown in Section 5.1. Further, given Assumption 7 ("large clusters"), the statistical error can be written as a closed-form function of m, as σ_k^2 approaches to zero ($k = 1 \ldots K$). In the following, the assumption of "small" volatilities is referred to as "Gaussian regime", because it implies $\Gamma_k \sim \mathcal{N}(1, \beta_k)$ ($k = 1 \ldots K$), as discussed in the proof of Theorem 3.

As in the previous Sections 3 and 4, both the standard CreditRisk$^+$ framework (Assumptions 1 and 3) and the modified "exponential" version (Assumptions 2 and 4) are discussed hereinafter.

In applications where c_h's are scarcely populated or σ_k's are not negligible, Theorem 3 is not guaranteed to cope with observations. This case is addressed in Section 5.2, where the robustness of the closed-form expression (54) is investigated by Monte Carlo simulations.

A numerical approach is maintained in Section 5.3 as well, where the estimation error of $\hat{A}$ at different time scales is measured in presence of autocorrelation, following the generalization introduced in Sections 3.3 and 4.4. In this case, the exponential version of the model comes in handy: indeed, it is observed that the error on the estimator introduced in (46) (i.e., standard CreditRisk$^+$ version) does not decrease at increasing m, while the opposite is true for the estimator presented in (49) (i.e., exponential CreditRisk$^+$ version).

In Section 4, the $\hat{A}$ estimator has been presented in multiple versions, depending on the considered model (standard or exponential version of CreditRisk$^+$), the chosen sampling period δ_m and the presence or absence of autocorrelation. Thus, it is worth introducing a compact notation to identify the different versions of $\hat{A}$.

The expressions for $A_{hh'}$ presented in (37) and (46) are addressed as "linear" estimators (as opposed to "exponential") in the following. In these cases the symbol $\hat{A}^{(L,m)}_{hh'}$ is used, where L stands for "linear" and $m = (T-t)/\delta_m$ is the ratio between the projection and calibration time scales.

On the other hand, the expressions for $A_{hh'}$ presented in (42) and (49) are addressed as "exponential" estimators and so the symbol $\hat{A}^{(E,m)}_{hh'}$ is used.

For the sake of brevity, when L or E is omitted, $\hat{A}^{(m)}_{hh'}$ refers to both the cases and, when m is omitted, $\hat{A}_{hh'}$ refers to the $m = 1$ case.

The improvement in statistical precision with respect to the estimate with no subsampling, can be quantified by the following ratio:

$$\varepsilon[\hat{A}^{(m)}_{hh'}] := \sqrt{\frac{\mathbf{var}\left[\hat{A}^{(m)}_{hh'}\right]}{\mathbf{var}\left[\hat{A}_{hh'}\right]}}. \tag{51}$$

Symbol $\varepsilon^{(m)}_{hh'}$ and its further specifications $\varepsilon^{(L,m)}_{hh'} := \varepsilon[\hat{A}^{(L,m)}_{hh'}]$ and $\varepsilon^{(E,m)}_{hh'}$ can be used as well. The last short notation that results to be convenient in the following is

$$c^{(Lm)}_{hh'} := \mathbf{cov}\left[F^{(j)}_h, F^{(j)}_{h'}\right] + s^{(j)}_h s^{(j)}_{h'}, \tag{52}$$

$$c^{(Em)}_{hh'} := \mathbf{cov}\left[L^{(j)}_h, L^{(j)}_{h'}\right]. \tag{53}$$

where $F^{(j)}_h$, $L^{(j)}_h$ and $s^{(j)}_h$ ($j = 1 \dots m$) are i.i.d. variables quantified using a sampling period δ_m.

Remark 5. *The notation "$\hat{A}$" refers to the fact the covariances involved in the definitions must be replaced with the corresponding sample estimators, when applying $A^{(m)}_{hh'}$ to historical time series. The same applies to the symbol $\hat{c}$.*

5.1. Precision of $\hat{A}$ at Different Time Scales under the Gaussian Regime

The following result quantifies the precision gain of performing CreditRisk$^+$ model calibration by historical time series available at increasing sampling frequencies. As anticipated, the precision of the estimated parameters increases as the sampling period decrease. This result holds under Assumption 7, in the limit $\sigma \to \mathbf{0}^+$ and considering absence of autocorrelation. The cases where some n_h is small (i.e., it does not verify Assumption 7) or where some σ_k is not negligible are addressed numerically in the next

Section 5.2—showing that the precision is still increasing as shorter sampling periods are considered. The introduction of autocorrelation is addressed in Section 5.3.

Theorem 3 (Estimation errors under Gaussian regime)**.** *Let us consider a set of risks, observed through a time interval* $(t_0, t]$ *and classified into a set of homogenous clusters* c_h $(h = 1 \dots H)$*. Let Assumption 3 ("CreditRisk*$^+$ *parameters at different time scales"), Assumption 5 ("non-autocorrelated market factors") and Assumption 7 ("Large clusters") hold with* $\xi_{kj} = 1$ *for a given uniform partition* $t_0 < t_1 < \cdots < t_j < \cdots < t_{m\times n} \equiv t$ *of* $(t_0, t]$*,* $(t_j - t_{j-1} = \delta_m;\ m, n \in \mathbb{N}^*)$*. Let* $\hat{A}$ *be the estimate of* A *needed to calibrate the CreditRisk*$^+$ *model in order to project losses over the time horizon* $(t, T]$*, such that* $(t - t_0)/(T - t) = n$ *and* $(T - t)/(\delta_m) = m$*. Then the following is true for* $\hat{A}^{(m)}_{hh'}$*:*

$$\varepsilon^{(m)}_{hh'} \xrightarrow{\sigma \to \mathbf{0}^+} \sqrt{\frac{n-1}{m \cdot n - 1}} \tag{54}$$

Equation (54) *remains true also considering Assumption 4 ("modified CreditRisk*$^+$ *parameters at different time scales") instead of Assumption 3.*

The proof of Theorem 3 is reported in Appendix A.5.

5.2. Beyond the Gaussian Regime: Numerical Simulations

In this section we verify that both the estimators $\hat{A}^{(E,m)}_{hh'}$ and $\hat{A}^{(L,m)}_{hh'}$ are more precise at increasing m. The closed-form results obtained in the Gaussian regime, discussed in Section 5.1, hold when the factor volatilities σ_Γ are much less than 1. Increasing σ_k $(k = 1, \dots, K)$ the Gaussian regime becomes less satisfactory and the difference of precision amongst determinations with different values of m becomes smaller. However, the error of $\hat{A}^{(m)}_{hh'}$ remains monotonically decreasing in m, even far from the Gaussian regime conditions.

We considered a case study with a two-factors market $(\Gamma_k,\ k = 1, 2)$. The couple of systematic factors induces the dependence between two populations of risks, as per the weights reported in Table 1.

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

k	**0**	**1**	**2**
ω_{1k}	0.30	0.40	0.30
ω_{2k}	0.50	0.25	0.25

The volatilities $(\sigma_k,\ k = 1, 2)$ associated to the factors are chosen according to seven different scenarios (indexed by i_σ), respectively as

$$\boldsymbol{\sigma}_\Gamma := 2^{i_\sigma} \begin{pmatrix} 2.5 \cdot 10^{-2} \\ 5.0 \cdot 10^{-2} \end{pmatrix}, \quad i_\sigma = 0 \dots 6. \tag{55}$$

For each scenario, the distributions of the estimators $\hat{A}^{(E,m)}_{12}$ and $\hat{A}^{(L,m)}_{12}$ $(m = 1 \dots 12)$ have been determined using 10^5 simulations 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. For both estimators the dynamic $\mathbf{F}_h(t, n_h)$ is that reported in (14). All risks belonging to the same cluster are supposed to have the same unconditioned intensity of default

$$q_i(t, t + \Delta_t) = -\frac{1}{\Delta_t} \log(0.99), \quad i = 1, \dots, n_h, \quad h = 1, 2. \tag{56}$$

To investigate the additional contribution to the error $\sigma\left[\hat{A}_{12}\right]$, generated by the finiteness of each cluster, different values of n_h have been considered. In particular, the number of claims

per each elementary temporal step $\delta_m = 1/m$ is extracted from a binomial distribution with parameter

$$n_h \in \left\{10^3, 2.5 \cdot 10^3, 5 \cdot 10^3, 10^4, 2.5 \cdot 10^4, 5 \cdot 10^4\right\}, \quad h = 1, 2. \tag{57}$$

For simplicity's sake, it is assumed that each defaulted risk is instantly replaced by a new risk, keeping the population of each cluster constant in time. Finally, the case $n_h = \infty$ (absence of binomial source of randomness) is also considered.

Figure 1 shows the behaviour of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ as a function of m, comparing various choices of σ_Γ. In this case we are not considering yet the contribution to error due to the finite population ($n_h = \infty$ for each cluster h). Equation (54) (red curve) is almost perfectly verified by the least volatility scenario ($\sigma_\Gamma = (2.5, 5.0) \cdot 10^{-2}$). At increasing volatility values (brighter curves), the gain in precision obtained at higher m is reduced, as well as the accordance with Equation (54).

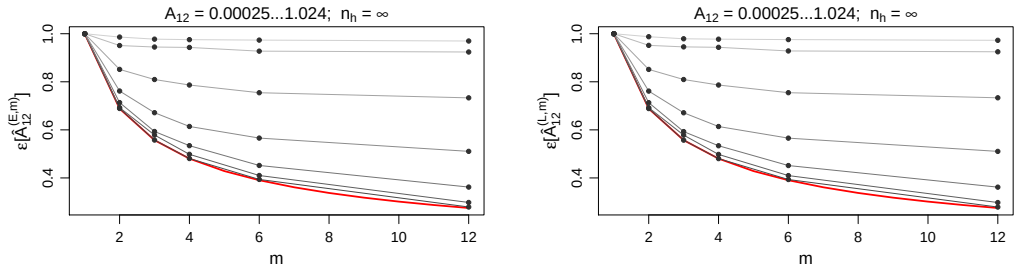

Figure 1. *Precision gain* $\varepsilon_{12}^{(m)}$*, as a function of* m *and* i_σ. The left and right plots show the values of $\varepsilon\left[\hat{A}_{12}^{(E,m)}\right]$ and $\varepsilon\left[\hat{A}_{12}^{(L,m)}\right]$ respectively, as a function of m, for each volatility scenario ($i_\sigma = 0, \ldots, 6$), each depicted with darker to brighter curves, in the $n_h = \infty$ assumption. The red curve is the theoretical value of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ in the Gaussian regime.

Since the transformation of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ moving away from the Gaussian regime (i.e., increasing $|\sigma_\Gamma|$) is smooth, estimating A_{12} with $m > 1$ remains convenient even for $[\sigma_\Gamma]_k \gtrsim 1$, despite the fact that Equation (54) is not verified anymore.

Comparison between the left and the right panel of Figure 1 shows that the above argument holds both in the linear and in the exponential case. This fact is also verified for all the other results of this section.

The results shown in Figure 1 are numerically checked against the case of finite portfolio populations: we tested each of the n_h declared in Equation (57). Even the smallest size considered (i.e., $n_h = 10^3$—Figure 2), that is affected by the largest binomial contribution to the error, leads to results comparable to the ones observed in the $n_h = \infty$ case. The size $n_h = 10^3$ is considered to be a limiting value for a realistic case.

We simulated the distribution of the estimator $\hat{A}_{12}^{(m)}$ as a function of m, testing all the possible combinations of σ_Γ and n_h declared in Equations (55) and (57). Figure 3 reports an example of the results. All the other considered (n_h, σ_k) couples resulted to have a similar behavior. The visual comparison between $\mathbf{E}[\hat{A}_{12}^{(m)}]$ (blue "X" symbol) and A_{12} level (red horizontal line) shows that indeed $\hat{A}_{12}^{(m)}$ is unbiased, both in the linear and in the exponential case (Equations (37) and (42) respectively). The dispersion around the mean reduces at increasing m, in agreement with both Equation (54) and the numerical results in Figures 1 and 2.

As implied by Figures 2 and 3, the number of risks n_h does not play a relevant role (if any) in computing the ratio $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$, while the absolute value of the standard error $\sigma\left[\hat{A}_{12}^{(m)}\right]$ is sensitive to the size of the portfolio.

This fact is confirmed by the results shown in Figure 4, where the estimates of $\sigma\left[\hat{A}_{12}^{(m)}\right]$ have been arranged as functions of n_h at fixed σ_Γ and m values. As expected, the standard error is greater when considering smaller n_h values, while the dependence on n_h of the error disappears quickly as approaching $n_h \to \infty$.

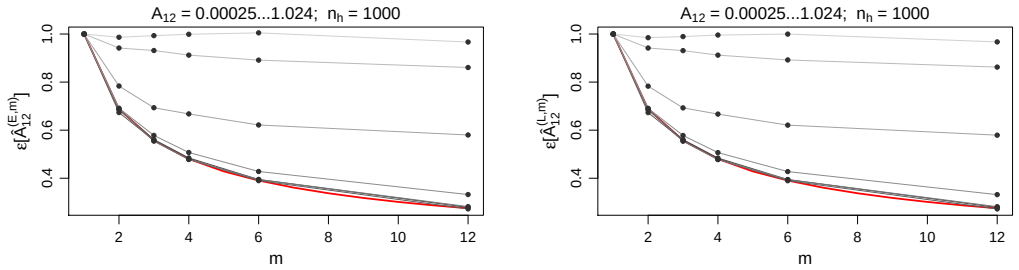

Figure 2. $\varepsilon[\hat{A}_{12}^{(E,m)}]$ *and* $\varepsilon[\hat{A}_{12}^{(L,m)}]$ *as a function of* m*, considering increasing* i_σ *(from darker to brighter curve) and* $n_h = 10^3$. The red curve is the theoretical value of $\varepsilon\left[\hat{A}_{12}^{(m)}\right]$ as a function of m in the 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 $\varepsilon\left[\hat{A}_{hh'}^{(L,m)}\right]$ is almost insensitive to n_h $(h = 1, 2)$.

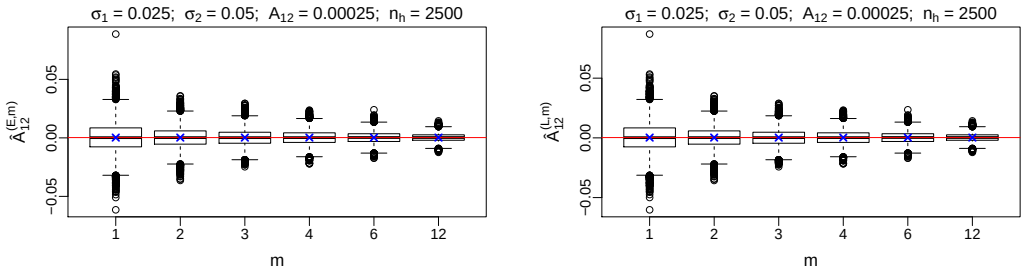

Figure 3. *Boxplot of* $\hat{A}_{12}^{(E,m)}$ *and* $\hat{A}_{12}^{(L,m)}$ *distributions, as a function of* m*.* The red horizontal line represent the true value of A_{12} and the blue X's stand for the average value of $\hat{A}_{12}^{(m)}$.

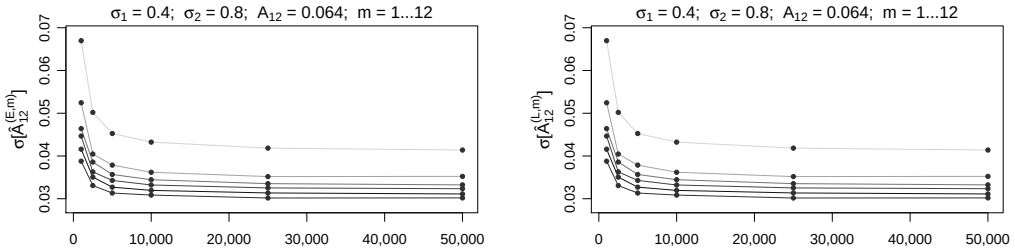

Figure 4. $\sigma[\hat{A}_{12}^{(E,m)}]$ *and* $\sigma[\hat{A}_{12}^{(L,m)}]$ *as a function of* n_h*.* Decreasing m values are considered from darker to brighter curve.

5.3. Estimation Error in Presence of Autocorrelation

In Section 5.2 the precision gain at increasing m is measured in absence of autocorrelation. In this section, the same numerical simulations are re-performed, introducing autocorrelation and comparing the results against the theoretical estimation of ε. The effect of autocorrelation on ε is discussed in Appendix B.

The numerical setup introduced above in Section 5.2 has been maintained, with a further assumption about ACF. Indeed, we assume that each latent variable $(k = 1, 2)$ obeys to the following ACF law, discussed in Section 3.3

$$\varrho_{xk} = \rho^{|x|}, \quad m = 12$$

where the considered ρ values are 0.05 and 0.5. For $m' < 12$ cases, we have considered ACF's resulting for the latent variables time series $\tilde{\Gamma}_k^{(j')}$ obtained by the clustering operation

$$\tilde{\Gamma}_k^{(j')} \equiv \frac{m'}{m} \sum_j \Gamma_k^{(j)}, \quad j = 1 + \tfrac{m}{m'}(j'-1) \dots \tfrac{m}{m'} j'$$

given the aforementioned ACF law at $1/m$ time scale. Since the contribution of the finite population to the error has been shown to be neglectable in Section 5.2, simulations in presence of autocorrelation have been performed under $n_h = \infty$ assumption only.

Figure 5 shows that the estimator $\hat{A}_{hh'}^{(E,m)}$ remains more precise at increasing m, even in presence of autocorrelation. The analytical results obtained in the Gaussian regime (i.e., theoretical superior and inferior estimates of ε—dashed and solid red lines in Figure 5), discussed in Appendix B, are in good agreement with the numerical results obtained in the considered set up. All the empirical measures of ε are included between the two theoretical limits (yellow areas).

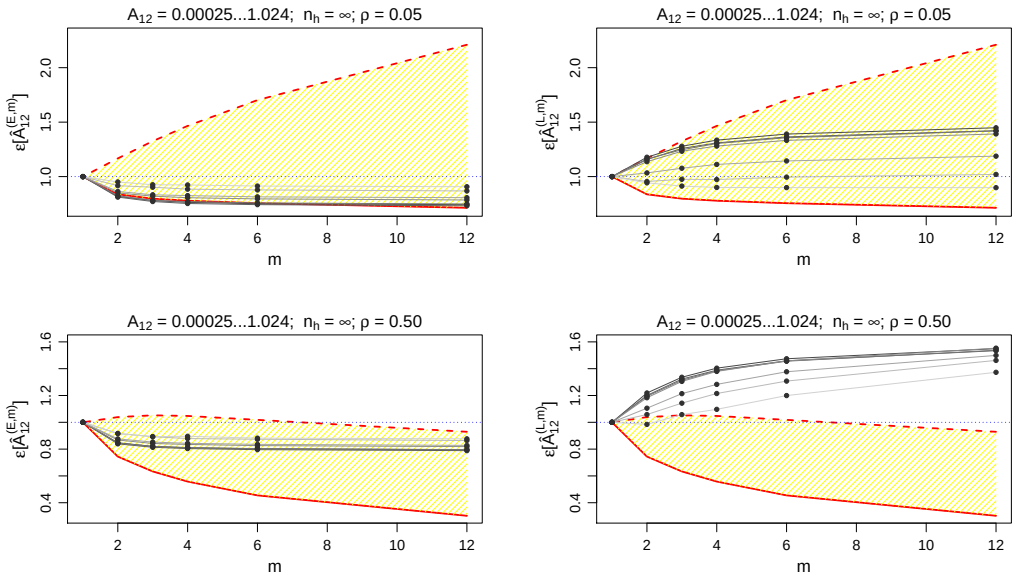

Figure 5. *Precision gain in presence of autocorrelation.* $\varepsilon[\hat{A}_{12}^{(E,m)}]$ (exact—**left** panels) and $\varepsilon[\hat{A}_{12}^{(L,m)}]$ (2nd order approximation—**right** panels), for each volatility scenario ($i_\sigma = 0, \dots, 6$, depicted with darker to brighter curves), for $\rho = 0.05$ (**top**) and $\rho = 0.5$ (**bottom**). The yellow area includes all the values between the maximum (dashed red line) and the minimum (solid red line) expected from the results of Appendix B. The frontier $\varepsilon = 1$ (dotted line) allows to check the presence of a precision gain at $m > 1$.

Moreover, precision gain (i.e., $\varepsilon < 1$) at $m > 1$ is also possible when using the estimator $\varepsilon\left[\hat{A}_{12}^{(L,m)}\right]$, introduced in Equation (46). However, due to the approximation introduced in this case, the estimator is not convenient (i.e., $\varepsilon > 1$) in the majority of the considered configurations.

6. An Application to Market Data

This section provides an example of the calibration technique applied to a real-world data set. The calibration technique is applied to a set of historical time series of bad loan rates supplied from the Bank of Italy. "Bad loan" is a subcategory of the broader class "Non-Performing Loan" and it is defined as exposures to debtors that are insolvent or in substantially similar circumstances [24].

In particular, the chosen data set is composed of the quarterly historical series TRI30496 ($m = 4$) over a five year period (from 1 January 2013 to 31 December 2017, $n = 5$, $\Delta_t = 1$).

The data are publicly available at [10]. The time series are supplied by the customer sector ("counterpart institutional sector") and geographic area ("registered office of the customer"). The latter, in the example, is held fixed to a unique value that corresponds to the whole country (Italy). Tables 2 and 3 report the definition of the 6 different clusters and their main features.

Table 2. Definition of the clusters $h = 1, \ldots, 6$ used in data set *TRI30496*.

Cluster Index h	Sector Code	Description
1	600	Consumer households
2	S11	Non-financial companies
3	S12BI7	Financial companies other than monetary financial institutions
4	S13	General government
5	S14BI4	Producer households
6	S15BI1	Non-profit institutions serving households and unclassifiable units

Table 3. Main features of the considered historical time series over the period 1 Gen 2013–31 Dec 2017. $\overline{p}_h$ ($h = 1, \ldots, 6$) is the yearly average bad loan rate; σ_h is the volatility associated to each $\overline{p}_h$; $\langle n_h \rangle$ is the average number of borrowers.

h	1	2	3	4	5	6
$\overline{p}_h$	0.0119	0.0352	0.0255	0.0056	0.0259	0.0088
σ_h	0.0010	0.0042	0.0023	0.0014	0.0022	0.0010
$\langle n_h \rangle$	269,515	407,602	3191	5416	132179	4020

By inspection of Table 3, it is possible to perform a rough estimate of σ_Γ. Equation (14) implies that the following holds for coefficients of variation CV_h ($h = 1, \ldots, H \equiv 6$):

$$CV_h := \frac{\sigma_h}{\overline{p}_h} \simeq \sum_{k=1}^{K} \omega_{hk} [\sigma_\Gamma]_k$$

Furthermore, the normalization requirement over the factor loadings ω_{hk} implies

$$\sum_{k=1}^{K} \omega_{hk} \lesssim 1$$

Hence we can state that $\langle CV \rangle := \frac{1}{H} \sum_h CV_h$ has the same order of magnitude of $\frac{1}{K} \sum_k [\sigma_\Gamma]_k$. Since $\langle CV \rangle \simeq 0.124$, results in Section 5.2 suggest that this data set is not far from the Gaussian regime and so there is an appreciable increase of precision in estimating $A(0,1)$ with $m > 1$.

$\hat{A}(0,1)$ is estimated by applying Equation (42) over a one-year period. The results obtained for $\hat{A}^{(E,m)}(0,1)$ ($m = 1, 4$) are reported in Table 4.

Table 4. Values of $\hat{A}^{(E,m)}(0,1)$ ($m = 4$ left, $m = 1$ right) obtained from the quarterly historical series *TRI30496* over the period 1 Gen 2013–31 Dec 2017. Results are expressed in 10^{-2} units.

0.53	0.28	0.33	0.36	0.41	0.48	0.68	0.40	0.36	1.01	0.56	0.73
0.28	0.59	0.48	0.61	0.43	0.40	0.40	1.50	0.98	1.26	0.87	−0.16
0.33	0.48	0.67	0.52	0.43	0.40	0.36	0.98	0.87	0.78	0.72	0.27
0.36	0.61	0.52	7.80	0.48	0.33	1.01	1.26	0.78	6.50	1.10	0.66
0.41	0.43	0.43	0.48	0.47	0.54	0.56	0.87	0.72	1.10	0.74	0.47
0.48	0.40	0.40	0.33	0.54	1.53	0.73	−0.16	0.27	0.66	0.47	1.35

The elementwise precision gain for $m = 4$, $\varepsilon\left[\hat{A}^{(E,4)}(0,1)\right]$, obtained under the Gaussian regime assumption, is shown in Table 5. This result is obtained applying definition (51) and Equation (A25) both to the cases $m = 4$ and $m = 1$. Equation (A25) has been shown to be valid under the Gaussian regime, discussed in Section 5.1.

In this case, the preliminary decomposition of $\hat{A}^{(E,4)}(0,1)$, that would be needed using the Monte Carlo method discussed in Section 5.2, is not needed.

Table 5. The elementwise precision gain $\varepsilon\left[\hat{A}^{(E,4)}(0,1)\right]$ associated with results reported in Table 4.

0.36	0.26	0.37	0.41	0.33	0.39
0.26	0.18	0.24	0.30	0.23	0.33
0.37	0.24	0.35	0.43	0.30	0.45
0.41	0.30	0.43	0.55	0.37	0.52
0.33	0.23	0.30	0.37	0.29	0.42
0.39	0.33	0.45	0.52	0.42	0.52

According to Equation (54), the elements of $\varepsilon\left[\hat{A}^{(E,4)}(0,1)\right]$ reported in Table 5 should be all approximately equal to 0.46, since they should depend only on the couple m, n ($m = 4$ and $n = 5$ in this case). However, in a real world case like the one considered, the assumption of zero autocorrelation is satisfied with a different precision by each time series $p_h(t)$. Furthermore, the estimated covariance matrices might need to be regularized (indeed the Higham regularization algorithm [25] was used both for $m = 1$ and $m = 4$ series). Hence, a different ratio for each element $(h, h') = 1, \ldots, 6$ is justified. Nonetheless, it is worth noticing that all the ratios reported in Table 5 have the same order of magnitude of the predicted value 0.46.

Knowledge of the historical number of risky subjects $n_h(t)$ for each cluster ($h = 1, \ldots, 6$) at each observation date ($t = 1/4, 2/4, \ldots, 5$) allows to take into account the binomial contribution to the error $\sigma\left[\hat{A}^{(E,m)}(0,1)\right]$, both for $m = 4$ (quarterly series) and $m = 1$ (yearly series), although the finiteness of the population does not add a relevant contribution to the error, as already observed in Section 5.2.

Table 6 provides Monte Carlo estimation of $\sigma\left[\hat{A}^{(E,m)}(0,1)\right]$ $(m = 1, 4)$, which considers also the role of $n_h(t)$. Since the values in Table 6 provide a measure of the error in the determination of $\hat{A}^{(E,m)}(0,1)$, it turns out that the estimates reported in Table 4 are elementwise consistent one with the other.

Table 6. $\sigma\left[\hat{A}^{(E,m)}(0,1)\right]$ ($m = 4$ left, $m = 1$ right). These are the elementwise errors of the estimators reported in Table 4. The results above are expressed in 10^{-2} units.

0.11	0.12	0.19	0.44	0.11	0.31	0.41	0.25	0.41	1.09	0.24	0.64
0.12	0.20	0.29	0.64	0.13	0.40	0.25	0.86	0.72	1.47	0.50	0.97
0.19	0.29	1.38	1.08	0.21	0.69	0.41	0.72	1.75	2.31	0.49	1.53
0.44	0.64	1.08	5.12	0.49	1.60	1.09	1.47	2.31	9.11	1.20	3.40
0.11	0.13	0.21	0.49	0.13	0.34	0.24	0.50	0.49	1.20	0.29	0.79
0.31	0.40	0.69	1.60	0.34	3.25	0.64	0.97	1.53	3.40	0.79	4.65

The Monte Carlo estimation of $\sigma\left[\hat{A}^{(E,m)}(0,1)\right]$, as done in Section 5.2, requires the a priori knowledge of the true dependence structure W, σ_Γ. Since this is a case study, we do not have an a priori parameterization of the calibrated model. Hence, we have used $\hat{W}, \hat{\sigma}_\Gamma$ estimated from $\hat{A}^{(E,4)}(0,1)$ instead, as a proxy of the "true" model parameters. The computation of $\hat{W}, \hat{\sigma}_\Gamma$ from $\hat{A}^{(E,4)}(0,1)$ is discussed below.

In order to complete the CreditRisk$^+$ calibration, we have to decompose $\hat{A}$ and find the factor loadings matrix $\hat{W}$ together with the vector of systematic factors variances $\hat{\sigma}_\Gamma^2$. To do so, we use the Symmetric Non-negative Matrix Factorization (SNMF), an iterative numerical method to search an approximate decomposition of $\hat{A}$ which satisfies the

requirements of the CreditRisk$^+$ model over $\hat{W}$ (i.e., all elements $\omega_{hk} > 0$ and $\sum_k \omega_{hk} = 1$). The application of SNMF to CreditRisk$^+$ is discussed in detail in [14]. In the following, we give evidence only of the implementation details necessary to address this case study. Being an iterative method, SNMF requires an initial choice of matrixes

$$\begin{aligned} \hat{U}_0 &:= \hat{W}_U \hat{\Sigma}^{1/2}, \\ \hat{V}_0 &:= \hat{\Sigma}^{1/2} \hat{W}_V, \end{aligned}$$

such that $\hat{A} = \hat{U}_0 \hat{V}_0$. It is not required that $\hat{U}_0 = \hat{V}_0^T$, nor all the elements of $\hat{U}_0$ and $\hat{V}_0$ have to be positive. We set $\hat{U}_0, \hat{V}_0$ from the eigenvalues decomposition of $\hat{A}^{(E,4)}(0,1)$.

For the considered data set, the eigenvalues decomposition returned the set of eigenvalues and eigenvectors reported in Table 7.

Table 7. Set of eigenvalues $\tilde{\sigma}_k$ and eigenvectors $\tilde{\omega}_k$ obtained by the eigenvalues decomposition of $\hat{A}^{(E,4)}(0,1)$, as reported in Table 4.

$\tilde{\omega}_k$	$k = 1$	$k = 2$	$k = 3$	$k = 4$	$k = 5$	$k = 6$
	0.06	−0.34	0.13	0.84	0.00	0.39
	0.10	−0.33	0.43	−0.38	−0.65	0.36
	0.09	−0.36	0.52	−0.26	0.72	0.06
	0.98	0.17	−0.07	0.01	0.01	0.00
	0.08	−0.38	0.22	0.19	−0.22	−0.84
	0.07	−0.68	−0.69	−0.20	0.06	0.07
$\tilde{\sigma}_k^2$	0.08	0.02	0.01	$2.9 \cdot 10^{-3}$	$1.4 \cdot 10^{-3}$	$0.3 \cdot 10^{-3}$

We use the $\tilde{\omega}, \tilde{\sigma}$ notation to address the quantities over which the normalization requirement of CreditRisk$^+$ has not been imposed yet.

Since more than the 95% of variance is explained by the first three eigenvectors, we reduced the dimensionality of the latent variables vector to be $K = 3$. Hence we define

$$\begin{aligned} \hat{U}_0 &= [\tilde{\omega}_1, \tilde{\omega}_2, \tilde{\omega}_3] \cdot \mathrm{diag}(\tilde{\sigma}_1, \tilde{\sigma}_2, \tilde{\sigma}_3) \\ &= \begin{bmatrix} 1.80 & 5.29 & 1.15 \\ 2.78 & 5.18 & 3.71 \\ 2.51 & 5.58 & 4.46 \\ 27.79 & -2.65 & -0.57 \\ 2.33 & 5.88 & 1.93 \\ 2.07 & 10.58 & -5.96 \end{bmatrix} \cdot 10^{-2} \end{aligned}$$

and $\hat{V}_0 = \hat{U}_0^T$. In general, SNMF aims to minimize iteratively the cost function

$$\left|\hat{A} - \hat{U}\hat{V}\right|^2 + \alpha \left|\hat{U} - \hat{V}^T\right|^2$$

where $|\cdot|$ is the Frobenius norm, eventually weighted, and α is a free parameter to weight the asymmetry penality term. Further details on the method are available in [14]. The application of SNMF method, together with the normalization constraint over the factor loadings, leads to the result reported in Table 8.

Table 8. The complete set of parameters $\hat{W}, \hat{\sigma}_\Gamma^2$ 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 4.

k	0	1	2	3
ω_{1k}	0.67	0.04	0.29	0.00
ω_{2k}	0.07	0.07	0.27	0.59
ω_{3k}	0.00	0.06	0.28	0.66
ω_{4k}	0.13	0.87	0.00	0.00
ω_{5k}	0.63	0.06	0.31	0.00
ω_{6k}	0.29	0.04	0.67	0.00
σ_k^2		0.103	0.031	0.010

A reasonable economic interpretation supports the set of parameters resulting from the calibration process described above. Indeed, factor loadings associated with the "general government" sector ($h = 4$) are completely distinct from the ones of the other sectors (i.e., this is the only sector mainly depending on the $k = 1$ factor): this fact copes with the different nature of the public entities from the ones belonging to the other considered sectors. Furthermore, "companies" ($h = 2, 3$) share approximately the same dependence structure. The same applies when considering "households" ($h = 1, 5$). Finally, the "institutions serving households" sector ($h = 6$) shares the same latent factor ($k = 2$) but shows a different balance between idiosyncratic and systematic factor loadings compared to "households", that is coherent with the nature of a sector strongly linked to "household" sectors, despite not being completely equivalent.

Results in Table 8 have been used to quantify the estimation errors reported in Table 6.

7. Conclusions

In this work, we have investigated how to calibrate the dependence structure of the CreditRisk$^+$ model, when the sampling period δ_m of the (available) default rate time series is different from Δ_t—the length of the future time interval chosen for the projections.

Preliminarily, we proved that CreditRisk$^+$ remains internally consistent when imposing the underlying distributional assumption to be simultaneously true at different time scales (Theorem 1). The model internal consistency is robust against the introduction of autocorrelation, depending on the considered ACF form (Theorem 2).

Then the problem has been approached in terms of moment matching, providing two asymptotically equivalent formulations for estimating the covariance matrix A amongst the systematic factors of the model (Propositions 2 and 3). The choice between the two estimators of A, provided in Equations (37) and (42), depends on the functional form (linear or exponential) that links the probability of claim/default and the latent variables. Both the estimators are explicitly dependent on the ratio Δ_t/δ_m, allowing for the calibration of the model at a time scale that is different from the one chosen for applying the calibrated model. Both the estimators have been generalized to autocorrelated time series in Equations (46) and (49), although only the latter (i.e., exponential case) is an exact result, while a second-order approximation has been adopted for the linear case.

Furthermore, calibrating the model on a shorter time scale than the projection horizon has been proved to be convenient in terms of reduced estimation error on $\hat{A}$. Analytical expressions for the error are provided in the Gaussian regime (i.e., small variances of the latent variables) by Theorem 3. In contrast, the case of increasing variance has been investigated numerically, confirming that, in general, the precision of the calibration is higher when employing historical data with a shorter sampling period. It has been verified that the convenience of calibrating the model at short time scales also remains in the presence of autocorrelation, although this is guaranteed only in the exponential framework, where an exact correction term is available.

Finally, the techniques presented in this work are shown to be numerically sound when applied to a real, publicly available data set of Italian bad loan rates.

Author Contributions: Conceptualization, J.G. and L.P.; methodology, J.G. and L.P.; software, J.G. and L.P.; validation, J.G. and L.P.; formal analysis, J.G. and L.P.; investigation, J.G. and L.P.; resources , J.G. and L.P.; data curation, J.G. and L.P.; writing—original draft preparation, J.G.; writing—review and editing, L.P.; visualization, J.G. and L.P.; supervision, J.G. and L.P.; project administration, J.G. and L.P.; funding acquisition, not applicable. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available in "Banca d'Italia—Base Dati Statistica" [10].

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Proofs

This section reports the proofs of theorems and propositions presented in this study.

Appendix A.1. Proof of Theorem 1

Proof. Firstly, the statement is proved considering Assumptions 1 and 3.

Assumption 3 implies by construction that $\{Y_i^{(j)}\}_{j=1\dots m}$ is a set of Poisson r.v.'s, which are mutually independent, conditionally on the realization of $\{\mathbf{\Gamma}^{(j)}\}_{j=1\dots m}$. Poisson distribution is closed with respect to addition. Hence

$$\sum_{j=1}^{m} Y_i^{(j)} | \mathbf{\Gamma}^{(j)} \sim \text{Poisson}(p_{i\Sigma}), \tag{A1}$$

where the distribution parameter is

$$p_{i\Sigma} = \sum_{j=1}^{m} \underbrace{q_i \frac{t_j - t_{j-1}}{T-t}}_{q_i^{(j)}} \left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik} \Gamma_k^{(j)} \right). \tag{A2}$$

Equation (20) in Assumption 3, the choice $\xi_{kj} = 1$ and the scaling property of Gamma distribution imply that

$$\frac{t_j - t_{j-1}}{T-t} \Gamma_k^{(j)} \sim \text{Gamma}\left(\sigma_k^{-2} \frac{t_j - t_{j-1}}{T-t}, \sigma_k^2 \right) \tag{A3}$$

Furthermore, Assumption 5 and the fact that independent Gamma r.v.'s with the same scale parameter are closed with respect to addition imply that

$$\sum_{j=1}^{m} \frac{t_j - t_{j-1}}{T-t} \Gamma_k^{(j)} \sim \text{Gamma}\left(\sigma_k^{-2}, \sigma_k^2 \right). \tag{A4}$$

Hence $\sum_{j=1}^{m} \frac{t_j - t_{j-1}}{T-t} \Gamma_k^{(j)} \equiv \Gamma_k$ and so $\sum_{j=1}^{m} Y_i^{(j)} \equiv Y_i$. This implies that $\{Y_i\}$ satisfies Assumption 1 over $(t, T]$.

The proof above can be extended to the exponential case, i.e., when considering Assumptions 2 and 4 instead of Assumptions 1 and 3. The form of parameter $p_{i\Sigma}$ in (A2)

can be obtained also can be obtained also from Assumption 4. In fact, the substitution $Y_i^{(j)} \mapsto \widetilde{Y}_i^{(j)}$ implies that $\widetilde{Y}_i \sim \text{Bernoulli}(\tilde{p}_i)$ where

$$\ln(1-\tilde{p}_i) = \ln \prod_{j=1}^{m} \left(1-\tilde{p}_i^{(j)}\right) = \sum_{j=1}^{m} q_i \frac{t_j - t_{j-1}}{T-t} \left(\omega_{i0} + \sum_{k=1}^{K} \omega_{ik} \Gamma_k^{(j)}\right). \tag{A5}$$

Considering Equation (A5) instead of (A2), the proof presented above holds for the $\widetilde{Y}_i$ representation of risks, ceteris paribus, implying that $\{\widetilde{Y}_i\}$ satisfies Assumption 2 over $(t, T]$. □

Appendix A.2. Proof of Theorem 2

Proof. The same arguments that lead to (A2) or to (A5) in proof of Theorem 1 are still valid in this case. Hence, it suffices to prove that mean and variance of the latent variable

$$\Gamma'_k := \sum_{j=1}^{m} \frac{t_j - t_{j-1}}{T-t} \Gamma_k^{(j)} = \frac{1}{m} \sum_{j=1}^{m} \Gamma_k^{(j)}$$

remain consistent with CreditRisk$^+$ requirements, stated in Assumption 1. It holds $\mathbf{E}[\Gamma'_k] = 1$, since $\mathbf{E}[\Gamma_k^{(j)}] = 1$. Moreover, the coefficient ξ_{jk} compensates the bias introduced in $\mathbf{var}[\Gamma'_k]$ by the fact that $\Gamma_k^{(j)}$ $(j = 1 \dots m)$ are autocorrelated according to the ACF ϱ_{xk}:

$$\begin{aligned} \mathbf{var}\left[\sum_{j=1}^{m} \Gamma_k^{(j)}\right] &= \sum_{j=1}^{m} \mathbf{var}\left[\Gamma_k^{(j)}\right] + \sum_{j=1}^{m} \sum_{j' \neq j} \mathbf{cov}\left[\Gamma_k^{(j)}, \Gamma_k^{(j')}\right] \\ &= \mathbf{var}\left[\Gamma_k^{(1)}\right] \underbrace{\left(m + 2 \sum_{x=1}^{m-1} (m-x) \varrho_{xk}\right)}_{m\xi_{kj}^{-2}} \end{aligned}$$

which implies $\mathbf{var}[\Gamma'_k] = \sigma_k^2$ directly.

The fact that Γ'_k is Gamma distributed is imposed in Assumption 6, implying that $\Gamma'_k \equiv \Gamma_k$ and so that Assumption 1 is satisfied. □

Appendix A.3. Proof of Proposition 2

Proof. Given a time interval $(t, T] \subseteq (t_a, t_b]$ and a uniform partition $(j = 1 \dots m)$ over $(t, T]$, Assumptions 3 and 5 imply that $\{Y_i\}$ satisfies Assumption 1 over $(t, T]$ by Theorem 1. Assumption 7 guarantees the convergence of F_h to $\mathbf{E}[F_h|\mathbf{\Gamma}]$ and of $F_h^{(j)}$ to $\mathbf{E}[F_h^{(j)}|\mathbf{\Gamma}^{(j)}]$, where we recall that $F_h = F_h(t, T)$.

For any interval $(t, T] \subseteq (t_a, t_b]$ and any pair of clusters $c_h, c_{h'}$, definitions (34), (35) and Assumption 5 imply that the covariance between F_{mh} and $F_{mh'}$ is given by

$$\mathbf{cov}(F_{mh}, F_{mh'}) = \prod_{j=1}^{m} \left[\mathbf{cov}\left(F_h^{(j)} F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)}\right] - s_h s_{h'} \tag{A6}$$

Since all the considered subintervals $(t_{j-1}, t_j]$ have the same length $\delta_m = t_j - t_{j-1}$, the frequencies $F_h^{(j)}$ are i.i.d., so that the above expression simplifies to:

$$\mathbf{cov}(F_{mh}, F_{mh'}) + s_h s_{h'} = \left[\mathbf{cov}\left(F_h^{(j)}, F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)}\right]^m \tag{A7}$$

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

Each cluster c_h is supposed to be homogenous by definition, i.e., $\boldsymbol{\omega}^{(i)} = \boldsymbol{\omega}^{(h)}$ for each risk $Y_i \in c_h$. Hence, distributional Assumptions 1 and 3 imply that both $F_h \xrightarrow{n_h(t)\to\infty} \mathbf{E}[F_h|\boldsymbol{\Gamma}]$ and $F_h^{(j)} \xrightarrow{n_h(t_j)\to\infty} \mathbf{E}[F_h^{(j)}|\boldsymbol{\Gamma}^{(j)}]$ are sample estimators of the parameters $p_h(\boldsymbol{\Gamma}) \coloneqq q_h(\omega_{h0} + \sum_k \omega_{hk}\Gamma_k)$ and $p_h^{(j)}(\boldsymbol{\Gamma}^{(j)})$ respectively, leading to the equivalence relation

$$F_{mh} = F_h = \hat{q}_h\left(\omega_{h0} + \sum_{k=1}^{K} \omega_{hk}\Gamma_k\right), \tag{A8}$$

therefore both $F_{mh}(t,T)$ and $F_h(t,T)$ are estimators of the default frequency for the $(t,T]$ interval. Thus, Equation (A7) can be rewritten as:

$$\mathbf{cov}(F_h, F_{h'}) + s_h s_{h'} = \left[\mathbf{cov}\left(F_h^{(j)}, F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)}\right]^m \tag{A9}$$

and, since $m = (T-t)/\delta_m$,

$$[\mathbf{cov}(F_h, F_{h'}) + s_h s_{h'}]^{1/(T-t)} = \left[\mathbf{cov}\left(F_h^{(j)}, F_{h'}^{(j)}\right) + s_h^{(j)} s_{h'}^{(j)}\right]^{1/\delta_m}. \tag{A10}$$

To complete the proof, let $(t,T]$ and $(t',T']$ be two subintervals of $(t_a,t_b]$, such that $(T-t)/(T'-t') \in \mathbb{Q}$. Hence, $\mathrm{GCD}\{T-t; T'-t'\} =: \overline{\delta} \in \mathbb{R}_+$ exists. $\overline{\delta}$ can be used as the mesh to define two uniform partitions over the two considered intervals.

Given these partitions, (A10) can be applied both to $T-t$ and to $T'-t'$, leading to

$$\begin{aligned}&[\mathbf{cov}(F_h(t,T), F_{h'}(t,T)) + s_h(t,T)s_{h'}(t,T)]^{1/(T-t)} = \\ &\left[\mathbf{cov}\left(F_h(t',T'), F_{h'}(t',T')\right) + s_h(t',T')s_{h'}(t',T')\right]^{1/(T'-t')}\end{aligned}$$

and completing the proof. The requirement $(T-t)/(T'-t') \in \mathbb{Q}$ can be easily weakened by the convergence of finite continued fractions with an increasing number of terms, until the desired degree of precision is reached. □

Appendix A.4. Proof of Proposition 3

Proof. Given a time interval $(t,T] \subseteq (t_a,t_b]$ and a uniform partition ($j = 1 \dots m$) over $(t,T]$, Assumptions 4 and 5 imply that $\{\widetilde{Y}_i\}$ satisfies Assumption 2 over $(t,T]$ by Theorem 1.

Assumption 7 guarantees the convergence of L_h to $\mathbf{E}[L_h|\boldsymbol{\Gamma}]$, where we recall that $L_h = L_h(t,T)$. Furthermore, it holds by definition that $\mathbf{E}[L_h|\boldsymbol{\Gamma}] = p_h(\boldsymbol{\Gamma})$, where the notation p_h has been introduced in the proof of Proposition 2.

The same apply to $L_h^{(j)}$ ($j = 1 \dots m$) for each uniform partition of $(t,T]$ considered; indeed, Assumption 7 implies $L_h^{(j)} \to \mathbf{E}[L_h^{(j)}|\boldsymbol{\Gamma}^{(j)}] = p_h^{(j)}(\boldsymbol{\Gamma})$.

Since $p_h(\boldsymbol{\Gamma}) = \sum_{j=1}^m p_h^{(j)}(\boldsymbol{\Gamma}^{(j)})$ and given that the partition is uniform, it holds $L_h = mL_h^{(j)}$ for each $j = 1 \dots m$. Since $m \coloneqq (T-t)/\delta_m$, we have

$$\frac{1}{T-t} L_h = \frac{1}{\delta_m} L_h^{(j)} \tag{A11}$$

Assumption 5 and Equation (A11) imply that

$$\frac{1}{T-t}\mathbf{cov}[L_h, L_{h'}] = \frac{1}{\delta_m}\mathbf{cov}[L_h^{(j)}, L_{h'}^{(j)}] \tag{A12}$$

for each considered pair of clusters $c_h, c_{h'}$. The proof is completed by the same argument used in proof of Proposition 2, after Equation (A10). □

Appendix A.5. Proof of Theorem 3

Proof. Assumptions 3 and 5 and $\xi_{kj} = 1$ imply Assumption 1 by Theorem 1. The same theorem implies Assumption 2 in case Assumption 4 is considered instead of Assumption 3, ceteris paribus. Furthermore, Assumptions 3, 5 and 7 and $\xi_{kj} = 1$ imply that

$$\hat{A}^{(L,m)}_{hh'} = \frac{1}{q_h q_{h'}}\left[\left(\hat{c}^{(Lm)}_{hh'}\right)^m - s_h s_{h'} - \delta_{hh'}\frac{q_h}{n_h}\right] \tag{A13}$$

by Proposition 2, for any $j = 1 \dots m$ and $h, h' = 1 \dots H$. Analogously, considering Assumption 4 instead of Assumption 3, it holds

$$\hat{A}^{(E,m)}_{hh'} = \frac{m}{q_h q_{h'}}\hat{c}^{(Em)}_{hh'} \tag{A14}$$

by Proposition 3, for any $j = 1 \dots m$ and $h, h' = 1 \dots H$.

The next step of the proof is showing that $\Gamma_k \tilde{\sim} \mathcal{N}(1, \beta_k)$ in the limit $\sigma_k \to 0^+$. In fact, both Assumptions 3 and 4 state that

$$\Gamma^{(j)}_k \sim \Gamma\left(\tfrac{1}{m\beta_k}, m\beta_k\right), \quad \mathbf{E}\left[\Gamma^{(j)}_k\right] = 1, \quad \mathbf{var}\left[\Gamma^{(j)}_k\right] = m\beta_k, \quad j = 1, \dots, m.$$

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

$$dF_k(x) \propto x^{(m\beta_k)^{-1}-1}\exp\left(-(m\beta_k)^{-1}x\right)dx \tag{A15}$$

Since it holds $(m\beta_k)^{-1} - 1 \xrightarrow{\sigma_k \to 0^+} (m\beta_k)^{-1}$, we have

$$\lim_{\sigma_k \to 0^+} dF_k(x) \propto \exp\left(\frac{\ln x - x}{m\beta_k}\right)dx. \tag{A16}$$

By introducing the auxiliary variable $x' := x - 1$ and replacing $\ln(1 + x')$ with the first three terms of its Maclaurin series, relation (A16) can be equivalently written as

$$\lim_{\sigma_k \to 0^+} dF_k\left(x(x')\right) \propto \exp\left(-\frac{x'^2}{2m\beta_k}\right)dx' \tag{A17}$$

In the limit $\sigma_k = \beta_k \to 0^+$, Equation (A17) implies that

$$\Gamma^{(j)}_k \sim \mathcal{N}\left(\mu = 1, \sigma^2 = m\beta_k\right). \tag{A18}$$

Hence it holds that each $F^{(j)}_h$ is normally distributed, with variance $m\sigma^2_h := m\sum_k \omega_{hk}\beta_k$—when considering the linear case (i.e., Assumptions 1 and 3). Analogously, also each $L^{(j)}_h$ is normally distributed in the exponential case (i.e., Assumptions 2 and 4).

Considering the market factors—as well as the historical observations of default frequency—as normal random variables is relevant to prove the theorem, since it implies that the covariance matrix estimators $\hat{c}^{(Lm)}$ and $\hat{c}^{(Em)}$ are Wishart distributed. Hence the variance associated to a given matrix element is

$$\mathbf{var}\left[\hat{c}^{(m)}_{hh'}\right] = \frac{m^2}{m \cdot n - 1}\left(\rho^2_{hh'} + 1\right)\sigma^2_h\sigma^2_{h'} \tag{A19}$$

in both linear and exponential cases. In the exponential case Equation (A19) is equivalent to the following

$$\mathbf{var}\left[\hat{c}^{(Em)}_{hh'}\right] = \frac{1}{m \cdot n - 1}\left[\left(c^{(Em)}_{hh'}\right)^2 + c^{(Em)}_{hh}c^{(Em)}_{h'h'}\right] \tag{A20}$$

while the same is not true in the linear case. Given Equation (A19), it is possible to prove Equation (54) separately in the two cases.

Proof in the linear case. Proposition 2 implies

$$\begin{aligned}\mathbf{var}\left[\hat{A}_{hh'}^{(L,m)}\right] &= \frac{1}{(q_h q_{h'})^2}\mathbf{var}\left[\left(\hat{c}_{hh'}^{(Lm)}\right)^m\right] = \frac{1}{(q_h q_{h'})^2}\left[\left(\mathbf{E}\left[\left(\hat{c}_{hh'}^{(Lm)}\right)^2\right]\right)^m - \left(\mathbf{E}\left[\hat{c}_{hh'}^{(Lm)}\right]\right)^{2m}\right] \\ &= \frac{1}{(q_h q_{h'})^2}\left[\left(\mathbf{var}\left[\hat{c}_{hh'}^{(Lm)}\right] + \left(\mathbf{E}\left[\hat{c}_{hh'}^{(Lm)}\right]\right)^2\right)^m - \left(\mathbf{E}\left[\hat{c}_{hh'}^{(Lm)}\right]\right)^{2m}\right] \end{aligned} \tag{A21}$$

In the limit $\sigma \to \mathbf{0}^+$ the binomial above can be replaced with its leading term. Hence

$$\mathbf{var}\left[\hat{A}_{hh'}^{(L,m)}\right] = \frac{1}{(q_h q_{h'})^2}\mathbf{var}\left[\hat{c}_{hh'}^{(Lm)}\right]\left(\mathbf{E}\left[\hat{c}_{hh'}^{(Lm)}\right]\right)^{2(m-1)} \tag{A22}$$

By applying Equation (A19) we have

$$\begin{aligned}\mathbf{var}\left[\hat{A}_{hh'}^{(L,m)}\right] &= \frac{1}{(q_h q_{h'})^2}\frac{m^2}{m\cdot n-1}\left(\rho_{hh'}^2+1\right)\sigma_h^2\sigma_{h'}^2\left(c_{hh'}^{(Lm)}\right)^{2(m-1)} \\ &= \frac{1}{(q_h q_{h'})^2}\frac{1}{m\cdot n-1}\frac{\rho_{hh'}^2+1}{\rho_{hh'}^2}\left(c_{hh'}^{(Lm)} - s_h^{(j)}s_{h'}^{(j)}\right)^2\left(c_{hh'}^{(Lm)}\right)^{2(m-1)} \end{aligned} \tag{A23}$$

Applying Proposition 2 once again we have $\left(c_{hh'}^{(Lm)}\right)^{2m} = \left(c_{hh'}^{(L1)}\right)^2$. Furthermore, we have $s_h^{(j)}s_{h'}^{(j)}\left(c_{hh'}^{(Lm)}\right)^{2(m-1)} \xrightarrow{\sigma\to\mathbf{0}^+} \left(s_h^{(j)}s_{h'}^{(j)}\right)^{2m} = s_h^2 s_{h'}^2$. Hence it holds

$$\mathbf{var}\left[\hat{A}_{hh'}^{(L,m)}\right] = \frac{1}{(q_h q_{h'})^2}\frac{1}{m\cdot n-1}\frac{\rho_{hh'}^2+1}{\rho_{hh'}^2}\left[\left(c_{hh'}^{(L1)}\right)^2 - s_h^2 s_{h'}^2\right] \tag{A24}$$

and thus the ratio $\mathbf{var}\left[\hat{A}_{hh'}^{(L,m)}\right]/\mathbf{var}\left[\hat{A}_{hh'}^{(L,1)}\right]$ verifies Equation (54), completing the proof for the linear case.

Proof in the exponential case. Equation (A20) and Proposition 3 imply

$$\begin{aligned}\mathbf{var}\left[\hat{A}_{hh'}^{(E,m)}\right] &= \frac{m^2}{(q_h q_{h'})^2}\mathbf{var}\left[\hat{c}_{hh'}^{(Em)}\right] = \frac{m^2}{(q_h q_{h'})^2}\frac{1}{m\cdot n-1}\left[\left(c_{hh'}^{(Em)}\right)^2 + c_{hh}^{(Em)}c_{h'h'}^{(Em)}\right] \\ &= \frac{1}{(q_h q_{h'})^2}\frac{1}{m\cdot n-1}\left[\left(c_{hh'}^{(E1)}\right)^2 + c_{hh}^{(E1)}c_{h'h'}^{(E1)}\right] \end{aligned} \tag{A25}$$

The latter implies that in case $m = 1$ we have

$$\mathbf{var}\left[\hat{A}_{hh'}^{(E,1)}\right] = \frac{1}{(q_h q_{h'})^2}\frac{1}{n-1}\left[\left(c_{hh'}^{(E1)}\right)^2 + c_{hh}^{(E1)}c_{h'h'}^{(E1)}\right] \tag{A26}$$

Hence, the ratio $\mathbf{var}\left[\hat{A}_{hh'}^{(E,m)}\right]/\mathbf{var}\left[\hat{A}_{hh'}^{(E,1)}\right]$ verifies Equation (54), completing the proof for the exponential case. □

Appendix B. Covariance Estimation Error in Presence of Autocorrelation

In this section a generalization of Equation (54) is provided, considering the presence of autocorrelation. Only the exponential case is discussed, because a closed form for $\hat{A}_{hh'}^{(E,m)}$ is still available when autocorrelation has to be considered—while only a second order approximation has been computed for the linear case $\hat{A}_{hh'}^{(L,m)}$.

A comparison between Equations (41) and (49) allows us to generalize Proposition 3.

$$c_{hh'}^{(E1)} = m c_{hh'}^{(Em)} + 2 \sum_{x=1}^{m-1} (m-x)\, {}_x c_{hh'}^{(Em)} \tag{A27}$$

where

$${}_x c_{hh'}^{(Em)} := \mathbf{cov}\left[L_h^{(j)}, L_{h'}^{(j+x)}\right] \tag{A28}$$

It holds by definition

$$\begin{aligned} c_{hh'}^{(Em)} &= \frac{q_h q_{h'}}{m^2} \sum_{k=1}^{K} \omega_{hk}\omega_{h'k} \mathbf{var}\left[\Gamma_k^{(j)}\right] \\ {}_x c_{hh'}^{(Em)} &= \frac{q_h q_{h'}}{m^2} \sum_{k=1}^{K} \omega_{hk}\omega_{h'k} \mathbf{cov}\left[\Gamma_k^{(j)}, \Gamma_k^{(j+x)}\right] \end{aligned}$$

Hence, Assumption 6 implies

$$\mathbf{E}\left[{}_x \hat{c}_{hh'}^{(Em)}\right] = \mathbf{E}\left[\frac{q_h q_{h'}}{m^2} \sum_{k=1}^{K} \omega_{hk}\omega_{h'k} \varrho_{xk} \mathbf{var}\left[\Gamma_k^{(j)}\right]\right] = {}_x \tilde{\varrho}_{hh'} \tag{A29}$$

where

$${}_x \tilde{\varrho}_{hh'} := \frac{\sum_{k=1}^{K} \tilde{w}_{khh'} \varrho_{xk}}{\sum_{k=1}^{K} \tilde{w}_{khh'}}; \quad \tilde{w}_{khh'} := \omega_{hk}\omega_{h'k} m \zeta_k^2 \sigma_k^2 \tag{A30}$$

Furthermore, applying Equation (A19), it follows that

$$\begin{aligned} \mathbf{var}\left[{}_x \hat{c}_{hh'}^{(Em)}\right] &= \frac{1}{m \cdot n - 1}\left(\mathbf{E}\left[{}_x \hat{c}_{hh'}^{(Em)}\right]^2 + \mathbf{E}\left[\hat{\mathbf{cov}}\left[L_h^{(j)}, L_h^{(j)}\right]\right]\mathbf{E}\left[\hat{\mathbf{cov}}\left[L_{h'}^{(j+x)}, L_{h'}^{(j+x)}\right]\right]\right) \\ &= \frac{1}{m \cdot n - 1}\left({}_x \tilde{\varrho}_{hh'}^2 \left(c_{hh'}^{(Em)}\right)^2 + c_{hh}^{(Em)} c_{h'h'}^{(Em)}\right) \\ &= \mathbf{var}\left[\hat{c}_{hh'}^{(Em)}\right] - \frac{1 - {}_x \tilde{\varrho}_{hh'}^2}{m \cdot n - 1}\left(c_{hh'}^{(Em)}\right)^2 \end{aligned} \tag{A31}$$

Equation (A29) leads to another version of Equation (A27)

$$c_{hh'}^{(Em)} = \tfrac{1}{m}\left(1 + 2 \sum_{x=1}^{m-1} (1 - \tfrac{x}{m})\, {}_x \tilde{\varrho}_{hh'}\right)^{-1} c_{hh'}^{(E1)} \tag{A32}$$

From Equation (49) we have

$$\mathbf{var}\left[\hat{A}_{hh'}^{(E,m)}\right] = \frac{m^2}{\left(q_h q_{h'}\right)^2} \mathbf{var}\left[\hat{c}_{hh'}^{(Em)} + 2 \sum_{x=1}^{m-1} (1 - \tfrac{x}{m})\, {}_x \hat{c}_{hh'}^{(Em)}\right] \tag{A33}$$

Equation (A33) implies that $\mathbf{var}\left[\hat{A}_{hh'}^{(E,m)}\right]$ depends on the correlation matrix $\varrho_{xx'}^{(\hat{c})}$ among the considered covariance estimators ${}_x \hat{c}_{hh'}^{(Em)}$ $(x = 0, 1, \dots)$, as shown below by choosing an equivalent expression for the RHS:

$$\mathbf{var}\left[\hat{A}_{hh'}^{(E,m)}\right] = \frac{m^2}{\left(q_h q_{h'}\right)^2} \sum_{x,x'=0}^{m-1} \varrho_{xx'}^{(\hat{c})}\, {}_x s_{hh'}\, {}_{x'} s_{hh'} \tag{A34}$$

where

$${}_x s_{hh'} := (2 - \delta_{0x})(1 - \tfrac{x}{m})\left(\mathbf{var}\left[{}_x \hat{c}_{hh'}^{(Em)}\right]\right)^{\frac{1}{2}} \tag{A35}$$

In case the covariance estimators are independent from each other (i.e. $\varrho^{(\hat{c})}_{xx'} = \delta_{xx'}$), an inferior limit to the considered variance is obtained

$$\mathbf{var}\left[\hat{A}^{(E,m)}_{hh'}\right] \geq \frac{m^2}{\left(q_h q_{h'}\right)^2}\left(\mathbf{var}\left[\hat{c}^{(Em)}_{hh'}\right] + 4\sum_{x=1}^{m-1}(1-\tfrac{x}{m})^2\mathbf{var}\left[{}_x\hat{c}^{(Em)}_{hh'}\right]\right) \tag{A36}$$

Equation (A31) can be substituted into Equation (A34). Hence, RHS of inequality (A36) becomes

$$\mathbf{var}\left[\hat{c}^{(Em)}_{hh'}\right]\frac{m^2}{\left(q_h q_{h'}\right)^2}\left[1 + 4\sum_{x=1}^{m-1}(1-\tfrac{x}{m})^2\left(1 - \tfrac{1-{}_x\tilde{\varrho}^2_{hh'}}{m\cdot n-1}\mathbf{cv}^{-2}\left[\hat{c}^{(Em)}_{hh'}\right]\right)\right]$$

where the notation $\mathbf{cv}[\cdot]$ stands for the coefficient of variation.

$(c^{(Em)}_{hh'})^2$ and $\mathbf{var}\left[\hat{c}^{(Em)}_{hh'}\right]$ can be expressed by using Equation (A32). Hence Equation (A36) can be used to estimate an inferior limit to $\varepsilon\left[\hat{A}^{(E,m)}_{hh'}\right]$ in the gaussian regime. A superior limit for the same quantity can be computed as well, imposing $\varrho^{(\hat{c})}_{xx'} = 1$ for each considered x, x'.

Remark A1. *Equation* (A34) *does not converge to* (A25) *in the limit* $\varrho_{xk} \to 0 \Rightarrow {}_x\tilde{\varrho}_{hh'} \to 0$. *This copes with the fact that assuming* $\varrho_{xk} = 0$ *in Equation* (A25) *implies a lesser error than measuring it.*

References

1. Crouhy, M.; Galai, D.; Mark, R. A comparative analysis of current credit risk models. *J. Bank. Financ.* **2000**, *24*, 59–117. [CrossRef]
2. Murphy, D. *Unravelling the Credit Crunch*; CRC Press: Boca Raton, FL, USA, 2009.
3. Schönbucher, P.J. *Credit Derivatives Pricing Models: Model, Pricing and Implementation*; Wiley: Hoboken, NJ, USA, 2003.
4. Fréchet, M. *Sur les Tableaux de Corrélation Dont les Marges Sont Donnés*; Annales de l'Université de Lyon, Science: Lyon, France, 1951; Volume 4, pp. 13–84.
5. Sklar, A. *Fonctions de Répartition à n Dimensions et Leurs Marges*; Institut Statistique de l'Université de Paris: Paris, France, 1951; Volume 8, pp. 229–231.
6. Joe, H. *Multivariate Models and Dependence Concepts*; Chapman and Hall/CRC: Boca Raton, FL, USA, 1997.
7. Nelsen, R.B. *Introduction to Copulas*, 1st ed.; Springer: Berlin/Heidelberg, Germany, 1999.
8. Li, D.X. On Default Correlation: A Copula Function Approach. *J. Fixed Income* **2000**, *9*, 43–54. [CrossRef]
9. Mai, J.-F.; Scherer, M. *Simulating Copulas: Stochastic Models, Sampling Algorithms, and Applications*, 2nd ed.; World Scientific Publishing Company: Singapore, 2017.
10. The Bad Loan Rate Series Is Labelled as *TRI30496_35120163*. The Count of Performing Borrowers at the Initial Period Series Is Labelled as *TRI30496_351122141*. Bank of Italy Statistical Database. Available online: https://infostat.bancaditalia.it/inquiry/ (accessed on 1 May 2021).
11. Credit Suisse First Boston. *CreditRisk$^+$, a Credit Risk Management Framework*; Credit Suisse First Boston: London, UK, 1998.
12. Passalacqua, L. *A Pricing Model for Credit Insurance*; Giornale Dell'Istituto Italiano Degli Attuari: Rome, Italy, 2006; Volume LXIX, pp. 1–37.
13. Passalacqua, L. *Measuring Effects of Excess-of-Loss Reinsurance on Credit Insurance Risk Capital*; Giornale Dell'Istituto Italiano Degli Attuari: Rome, Italy, 2006; Volume LXX, pp. 81–102.
14. Vandendorpe, A.; Ho, N.D.; Vanduffel, S.; Van Dooren, P. On the parameterization of the CreditRisk$^+$ model for estimating credit portfolio risk. *Insur. Math. Econ.* **2008**, *42*, 736–745. [CrossRef]
15. Wilde, T. CreditRisk$^+$. In *Encyclopedia of Quantitative Finance*; John Wiley & Sons: Hoboken, NJ, USA, 2010.
16. Gundlach, M.; Lehrbass, F. (Eds.) *CreditRisk$^+$ in the Banking Industry*; Springer: Berlin/Heidelberg, Germany, 2004.
17. Klugman, S.A.; Panjer, H.H.; Willmot, G.E. *Loss Models: From Data to Decisions*; Wiley: Hoboken, NJ, USA, 2012.
18. Glasserman, P.; Li, J. Importance Sampling for Portfolio Credit Risk. *Manag. Sci.* **2005**, *51*, 1643–1656. [CrossRef]
19. McNeil, A.; Frey, R.; Embrechts, P. *Quantitative Risk Management*; Princeton University Press: Princeton, NJ, USA, 2015.
20. Kotz, S.; Adams, J.W. Distribution of Sum of Identically Distributed Exponentially Correlated Gamma-Variables. *Ann. Math. Stat.* **1964**, *35*, 277–283. [CrossRef]
21. Mathai, A.M.; Moschopoulos, P.G. A Form of Multivariate Gamma Distribution. *Ann. Inst. Stat. Math.* **1992**, *44*, 97–106. [CrossRef]
22. Florent, C.; Borgnat, P.; Tourneret, J.; Abry, P. Parameter estimation for sums of correlated gamma random variables. Application to anomaly detection in Internet Traffic. In Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP-08, Las Vegas, NV, USA, 31 March–4 April 2008.

23. Feng, Y.; Wen, M.; Zhang, J.; Ji, F.; Ning, G. Sum of arbitrarily correlated Gamma random variables with unequal parameters and its application in wireless communications. In Proceedings of the IEEE 2016 International Conference on Computing, Networking and Communications (ICNC), Kauai, HI, USA, 15–18 February 2016. [CrossRef]
24. Non-Performing Loans (NPLs) in Italy's Banking System. 2017. Available online: https://www.bancaditalia.it/media/views/2017/npl/ (accessed on 1 May 2021).
25. Higham, N. Computing the nearest correlation matrix—A problem from finance. *IMA J. Numer. Anal.* **2002**, *22*, 329–343. [CrossRef]