Research Article

Enrico Bernardi and Silvia Romagnoli

A copula-based hierarchical hybrid loss distribution

Abstract: We propose a model for the computation of the loss probability distribution allowing to take into account the not-exchangeable behavior of a portfolio clustered into several classes of homogeneous loans. These classes are classified as 'large' or 'small' depending on their cardinality. The hierarchical hybrid copula-based model (HHC for short) follows the idea of the clusterized homogeneous copula-based approach (CHC) and its limiting version or the limiting clusterized copula-based model (LCC) proposed in our earlier work. This model allows us to recover a possible risk hierarchy. We suggest an algorithm to compute the HHC loss distribution and we compare this cdf with that computed through the CHC and LCC approaches (in the Gaussian and Archimedean limit) and also with the pure limiting approaches which are commonly used for high-dimensional problems. We study the scalability of the algorithm.

Keywords: Hierarchical copula functions, limiting loss distribution, clusterized copula function

MSC 2010: 62H20, 60C05, 90B50, 91G40

DOI: 10.1515/strm-2012-1128
Received March 27, 2012; revised February 13, 2015; accepted March 18, 2015

1 Introduction

Copula functions (see [18, 21]) represent a methodology which has recently become the most significant new tool to describe the co-movement between markets, risk factors and other relevant variables studied in finance. This useful technique is applied to several applied mathematical fields (see [6] for standard applications of copula in finance, and [5] for the new frontiers of dynamic copula methods in finance). For the application of copula functions to credit risk modeling we mention [13, 19, 25]. Since our interest lies in proposing a new copula-based methodology in order to compute the distribution of losses, we recall that in the pure limiting Gaussian model (PL Gaussian for short) as suggested by Vasicek [28, 29] the cdf of losses of a large portfolio was described via the inverse Gaussian distribution function, whose simple closed-form solution was quickly adopted in practice. In the same spirit, using an algorithm from the theory of the Archimedean copula function, Schönbucher [24] gave some more limiting loss distributions, which are driven by random variables with different dependency structures. We observe that this pure limiting distribution (PL Archimedean for short) may not be useful for portfolio credit risk clustered into classes affected by different risk factors, given the exchangeability property of the Archimedean copula.

Along the same lines of [17, 23], here we propose a new model to compute the loss distribution, which accounts for risk hierarchy. Choroś-Tomczyk et al. [9] proposed a very flexible model based on hierarchical Archimedean copulas to value CDOs allowing for non-exchangeable dependency structures. In the same way the suggested model is a hierarchical approach since it considers more than one level of risk and can be considered as an hybrid version of the CHC and the LCC models in [3, 4]. Therefore we point out that the model is hybrid, that is, at the lowest level of risk, it selects when to determine the loss pdf through a limiting model or rather through a clusterized one, depending on the cardinality of the risk class. In more detail, we consider the default event, and we recover the cumulative probability distribution of the variable counting

Enrico Bernardi, Silvia Romagnoli: Department of Statistics, University of Bologna, Via Belle Arti 41, 40126 Bologna, Italy, e-mail: enrico.bernardi@unibo.it, silvia.romagnoli@unibo.it

the number of events[1] occurred, assuming a portfolio clustered into n sub-portfolios of homogeneous loans, meaning that the loans of the same class are affected by the same risk factor. Once the loss distribution into each class is computed by the selected method (limiting or CHC), the aggregation between classes is derived by considering the dependence structure at the highest level. Our problem is strictly linked with the literature on concentration risk composed by a granularity risk and a sector concentration risk. Granularity risk is not considered into the *large* sub-portfolio, where we state the infinite granularity assumption, but is considered into the *small* ones; anyway a semi-granularity adjustment (see [15, 16]) may be also considered into the 'within groups' setting. On the other hand, here we are interested in the sector concentration risk in the same spirit of the multi-factor models (see [22, 27]). We focus on a clustered portfolio highlighting another component of the sector concentration which comes from the assumption on the affecting policy; if we assume a non-equally affected policy but we consider all the possible combinatorial distributions of the percentage of defaults into the groups, it is then possible to disentangle the impact of a particular concentration scenario as a relevant factor in order to determine the expected loss distribution.

Following the approach of [3, 4], we focus on the problem to recover the probability distribution of the counting variable linked to the copula function representing the dependence between the classes, and we suggest an algorithm to implement it. We do not address here the problems concerning the construction of the clustered copula function, i.e. the clustering problem concerning the choice of the metric and of the number of clusters and the homogenization problem, for which we refer to [2]. Nevertheless, we point out that the necessary reduction of the complexity of an high-dimensional problem, in order to make it tractable, can be reached through a clusterization into *equivalent* homogeneous classes; for this reason copulas based on pair-copula constructions, as vine copulas (see [1]), are not suitable to our aim.

The paper is organized as follows: In Section 2, we present and discuss the HHC approach concerning the *within* and the *between* groups dependence structure to recover the probability distribution of the first level counting variable. In Section 3, we propose some experiments starting from four basic concentration scenarios of a credit-exposed portfolio. A comparative analysis with respect to the loss cdf computed with the PL and LCC Gaussian and Archimedean approaches is outlined. The scalability and the precision of the algorithm implementing the HHC model is analyzed.

2 The hierarchical hybrid copula-based approach

We consider a portfolio credit risk model where we assume that the basket of loans is clustered into a finite number n of sub-portfolios or classes affected by different risk factors. The basic idea is to introduce a hierarchical risk structure allowing to take into account the non-exchangeability of the variables and the grains of such classes. The proposed HHC model is composed by three steps: a preparing phase, a *within class* computing step and a *between classes* aggregation step.

2.1 The preparing phase

The starting implementation of the proposed methodology is concerned basically with the reduction of the complexity of the problem given the dimension of the data set. This aim may be reached by a clusterization and a homogenization procedure.

We consider a set of data and we assume to partition the given data into so-called homogeneous clusters such that patterns within a cluster are more similar to each other than patterns belonging to different clusters. The best known and most widely used algorithm to reach this goal is K-means, which is a semi-unsupervised approach, and where one needs to choose the distance metric and the number of clusters. Francois et al. [12] pointed out that all metrics are substantially equivalent in a not randomized data set while the crucial vari-

[1] We count a number of a defined percentage unit of a variable (losses) attached by the event (of default).

able is the number of clusters. The choice of this variable involves the consideration of the trade-off between the complexity of the problem in term of time of computing and the precision of the method. In particular, a greater number of groups probably means a greater number of *small* classes and then a lower error of granularity; the implied *whole loss cdf* will be then more conservative as we can observe in the experiments proposed in Section 3. On the other hand, the unsupervised methods, like the SOM approach, choose automatically the best number of clusters to made. We refer to [2] for a discussion on these choices.

In the experiments proposed in Section 3, we opted for K-means with the Euclidean metric and with a number of clusters optimal in some sense, i.e. for example minimizing a clustering error defined as a sum of *confusions* (see [30]). The method proposed here, however, may be implemented by similarly grouping the data with other clustering methods. Anyway, for a *good* choice of the number of clusters, the performance of the semi-unsupervised approaches is comparable with that of the unsupervised methods implying that it does not substantially affect the final loss cdf. Nevertheless, the unsupervised methods are more stable than the semi-unsupervised ones in the randomized data sets.

The clustering method assembles some groups, which must be considered homogenous to simplify the computation in the case of high-dimensional problems. To this aim we follow a procedure along the lines of the diversity score approach suggested in [7, 8] and applied in [11, 26]. In order to compensate the homogenous assumption with a change in the number of variables inside the group, we assume the cardinalities of the groups of the clusterized copula function to be *equivalent* in the meaning adopted in [2].

2.2 The ‘within classes’ computing step

We consider a portfolio clustered into a finite number of classes and compute the *within classes* loss distributions. If the class is *large*[2], we compute the corresponding loss distribution by a limiting approach like the Gaussian or the Archimedean ones. On the other hand, if the class is *small*, we recover the loss pdf through the CHC approach.

2.2.1 Large classes

We assume that the j-th *large* class is characterized by the following assumption:

- The class is composed by I_j obligors having the same exposure size and the same loss in default. The number of obligors I_j is very large implying that the relevant quantity for the portfolio risk is the fraction L^j of defaulted obligors in the j-th sub-portfolio.
- All obligors have the same unconditional default probability $p_i^j = p^j$ for all i until a fixed-horizon time T.

These assumptions justify the application of a PL model to compute the j-th loss cdf.

The Gaussian PL model suggested in [28, 29] proposes the following loss cdf within the j-th large group:

$$F_j^l(q) = \mathbb{P}(L^j \le q) = \Phi\bigg(\frac{1}{\sqrt{\rho_j}}\Big(\sqrt{1-\rho_j}\,\Phi^{-1}(q) - \Phi^{-1}(p^j)\Big)\bigg), \tag{1}$$

where p^j is the default probability of any individual obligor in the j-th class, $\Phi(\cdot)$ denotes the cumulative standard normal distribution function, and ρ_j is the asset value correlation between any two obligors of j-th class.

In the limiting Archimedean approach proposed in [24], the loss distribution of the j-th class can be derived by conditioning on the mixing variable Y^j representing the common risk factor affecting the j-th loans' group (see [20]).

[2] We say that the class is *large* if its cardinality is greater than or equal to a fixed bound. Here we justify the assumption of *large* portfolio if the group cardinality is greater than the threshold of 20, as suggested in [24].

The probability of having not more than a fraction q of defaults in the j-th portfolio is

$$F_j^l(q) := \mathbb{P}(L^j \leq q) = 1 - G_j\Big(-\frac{\log q}{\phi_j(p^j)}\Big), \tag{2}$$

where $G_j(\cdot)$ is the Y^j's cdf and the limiting loss pdf of the j-th group is

$$f_j^l(q) = \frac{1}{q\phi_j(p^j)} g_j\Big(-\frac{\log q}{\phi_j(p^j)}\Big),$$

where g_j is the pdf of the mixing variable.

2.2.2 Small classes

If the class is *small*, we compute the relative loss pdf through the *clusterized homogeneous copula* approach, CHC for short (see [3]). We review in the following the essentials about this approach which allows us to extract the loss pdf from the dependence structure of the elements of a *small* class. This dependence structure is modeled through a CHC that provides a second level of clusterization inside the *small* groups. CHC approach can be seen as a correction of the limiting models for granularity and concentration risk. Granularity risk is considered into the CHC model since each group is composed by a finite number of homogeneous variables while the sector concentration risk is taken into account through the clusterization of the variables. In [3], we compared the performance of the CHC model for several numbers of groups and showed that it is better than that of other models. Clearly this model loses a part of information of the true distribution to reduce the complexity of the problem, but, as expected, the losing information decreases if the number of groups increases. We can observe that a CHC with unit groups (U-CHC for short) is able to recover the *exact* loss cdf and to completely correct for the granularity error. The comparison between a U-CHC loss cdf and a limiting loss cdf properly gives the dimension of the granularity error while the comparison between a whole U-CHC and the U-HHC, i.e. an HHC whose small classes are U-CHC, gives the *semi-error of granularity* since it is an error affecting only large classes. On the other hand, the comparison between a U-HHC and a U-HHC in a concentration scenario where all classes are large defines the so-called *partial error of granularity*. We analyze the impact of these errors in several scenarios experimented in Section 3.

CHC is a hierarchical copula function where we have only one level of dependence, i.e. where the dependence parameters between the groups are equal to the dependence parameters within the groups. It is useful to reduce the complexity of the problem also if we have one level dependence structure. In [2], a volume's evaluation approach is proposed for a multilevel dependence structure, i.e. for a proper hierarchical copula function.[3]

We assume that the j-th *small* class is composed by n_j elements clusterized into k_j homogeneous groups of *equivalent*[4] dimension m_s^j, $s = 1, \dots, k_j$, respectively, such that $\sum_{s=1}^{k_j} m_s^j = n_j$. The most important ingredient to extract the loss pdf from a CHC is the volume of this kind of copula. We recall the representation of this volume proposed in [2] for homogeneous groups.

[3] We observe that a proper hierarchical structure affects the distribution of the counting variable which could not be recovered through the combinatorial algorithm training the CHC approach. In this case, we should associate to the compatible combinatorial distributions a probabilistic distribution, allowing to consider the policy of affecting the groups implied by the hierarchy. The hierarchy imposes a kind of weighting function to the compatible combinatorial distributions transforming the associated counting variable into an object that is no longer a proper random variable. To determine the probability distribution of this not properly counting variable linked to a hierarchical structure, we need a new probabilistic instruments. This is a promising field for future research.

[4] Here equivalent is used in the meaning of the diversity score approach (see [7, 8]).

Proposition 2.1 (CHC$_j$ volume). *Given a* CHC$_j$ *representing the dependence structure of the j-th class, the volume of the k_j-dimensional box* $\mathbf{S} = [\mathbf{u}, \mathbf{v}]$ *with* $\mathbf{u}, \mathbf{v} \in [0,1]^{k_j}$, $\mathbf{u} \leq \mathbf{v}$, *may be represented as:*

$$V_{\mathrm{CHC}_j}(\mathbf{S}) = \sum_{i=0}^{n_j} (-1)^i \sum_{z=1}^{\hat{D}^c(i,k_j)} \mathrm{CHC}_j(\mathbf{c}(\mathbf{p}_{i,k_j}(z))),$$

where

- $\hat{D}^c(i, k_j)$ *denotes the number of compatible combinatorial distributions (c.c.d. for short)*[5] *of the integer i into k_j groups;*
- $\mathbf{p}_{i,k_j}(z)$ *is a* $k_j \times \max(m_s^j, s = 1, \dots, k_j)$ *matrix of zeros and ones corresponding to the z-th c.c.d. of i ones into k_j groups of dimensions* m_s^j, $s = 1, \dots, k_j$;
- $\mathbf{c}(\mathbf{p}_{i,k_j}(z))$ *is a* $k \times \max(m_s^j, s = 1, \dots, k_j)$ *matrix such that* $c_{s,w} = v_s$ *if* $p_{s,w} = 0$ *and* $c_{s,w} = u_s$ *if* $p_{s,w} = 1$, *where* u_s, v_s *is the s-th element of* $\mathbf{u}, \mathbf{v}$, *respectively, and* $c_{s,w}$ *denotes the* (s, w)*-th element of the corresponding matrix;*
- $\mathrm{CHC}_j(\mathbf{c}(\mathbf{p}_{i,k_j}(z)))$ *is the clusterized copula computed for the z-th c.c.d. accounting for the permutation of the elements into the groups.*

As pointed out before, we are interested in an event which may involve a basket of variables, composing a *small* class, and in defining a counting variable on this event that we call *second level counting variable*[6]. Here we assume to count some unit of percentage losses. We observe that the *second level counting variable* counts the number of unit percentage losses inside the j-th *small* class and then corresponds to the variable L^j yet introduced for the *large* classes.

Definition 2.2 (Second level counting variable linked to a CHC$_j$). In the same setting of Proposition 2.1, the counting variable linked to the $\mathrm{CHC}_j(\mathbf{c}(\mathbf{p}_{l,k_j}(z)))$ is given by $r(z) = \sharp(\mathbf{p}_{l,k_j}(z))$ which counts the number of elements equal to one in the matrix $\mathbf{p}_{l,k_j}(z)$ for the z-th c.c.d. Clearly we have $r(z) = l$ for all z.

Our aim is to extract the loss pdf of the j-th *small* class from the CHC$_j$ representing the dependence structure of the j-th class. To achieve this, we observe that a number l corresponds, through its probability function, to a number in $[0, 1]$ by considering all the c.c.d. of l ones into k_j groups (we enumerate this c.c.d. with the index z). The z-th c.c.d., corresponding to the matrix $\mathbf{p}_{l,k_j}(z)$, generates a pair of coordinates[7], i.e. the z-th c.c.d. generates the box $\mathbf{S}^z$ for which we compute the volume. The sum of the volumes computed for all the coordinates generated by all the c.c.d. for the same number of ones, i.e. for the same l ones into k_j groups, represents the probability to count l ones, or better to have l events whose probabilities are the marginal centroids of the groups.

Proposition 2.3 (Loss cdf linked to the CHC$_j$). *In the same setting of Proposition 2.1, the loss cdf linked to the* CHC$_j$, *with centroids* $\mathbf{u} \in [0, 1]^{k_j}$, *is the function* $F_j^s : [0, n_j] \to [0, 1]$ *such that*

$$F_j^s(i) = \sum_{l=0}^{i} \sum_{z=1}^{\hat{D}^c(l,k_j)} V_{\mathrm{CHC}_j}(\mathbf{S}^z(\mathbf{p}_{l,k_j}(z)))$$

where $V_{\mathrm{CHC}_j}(\mathbf{S}^z)$ *is the volume of the* CHC$_j$ *computed for the box* $\mathbf{S}^z = [\mathbf{u}^z, \mathbf{v}^z] \in \mathbb{R}^a \times \mathbb{R}^a$, $\max(a) = 2k_j$, *determined for the z-th c.c.d. of l ones into k_j groups, where* $\mathbf{p}_{l,k_j}(z)$ *is a* $k_j \times \max(m_s^j, s = 1, \dots, k_j)$ *matrix of zeros and ones corresponding to the z-th c.c.d. of l ones into k_j groups of dimensions* m_s^j, $s = 1, \dots, k_j$, *and finally accounting for the permutation of the elements into the groups.*

[5] We refer to $\hat{D}(i, k_j)$ as the number of the ways in which one can distribute the integer i into k_j groups without taking into account the order of the groups (c.d. for short). If we take into account the equivalent cardinalities of the groups, then we refer to $\hat{D}^c(i, k_j)$ as the number of the compatible (arranged for the group cardinalities) combinatorial distributions.

[6] This is a *second level counting variable* since it is linked to a second level dependence structure or a dependence structure within a class.

[7] The rule which explains how the coordinates are generated is presented in [3].

2.3 The 'between classes' aggregation step

We are interested in determining the distribution of the whole loss defined as the sum of the classes' loss, $L = \sum_{j=1}^{n} L^j$, where $n = n_l + n_s$.

Now the problem is to associate a counting variable to the copula function representing the dependence structure between the classes; this *first level counting variable*, which represents the *whole loss cdf*, distributes the losses to the classes.

Given the vector of the loss cdf within classes, with a threshold of 20 as proposed in [24][8],

$$\mathbf{F}(\mathbf{q}) = \big(F_j^{\varepsilon}(q_j),\ j = 1, \dots, n_l + n_s,\ (\varepsilon = s)\mathbf{1}_{n_j<20} \cup (\varepsilon = l)\mathbf{1}_{n_j \geq 20}\big),$$

composed by limiting loss cdf for *large* classes and by CHC-based loss cdf for *small* classes, we are interested in aggregating these functions with a dependence structure. By Sklar's theorem, these functions being uniformly distributed, we can represent their dependence structure as a copula $C(F_1^{\varepsilon}(q_1), \dots, F_n^{\varepsilon}(q_n))$, where $n = n_l + n_s$. This is a hierarchical structure characterizing the highest or the first level of dependence.

We consider the default event and the corresponding counting variable on this event, that we call *first level counting variable*. Here we assume to count some unit of percentage losses.

Definition 2.4 (First level counting variable linked to an n-dimensional copula function). The first level counting variable linked to the $C(\mathbf{F}(\mathbf{p}_{l,n}(j)))$, where $\mathbf{p}_{l,n}(j)$ is the j-th c.c.d.[9] of l units percentage losses into n groups and $\mathbf{F}(\mathbf{p}_{l,n}(j))$ is the corresponding vector of margins computed with a limiting or CHC-based approach depending on the cardinalities of the groups, is given by

$$wr(j) = \sharp(\mathbf{p}_{l,n}(j))$$

which counts the elements of vector $\mathbf{p}_{l,n}(j)$ for the j-th c.c.d. Clearly we have $wr(j) = l$ for all j.

Our aim is to recover the *whole loss cdf* or the cdf of the *first level counting variable* linked to the n-dimensional copula function C. To do this we observe that a number l corresponds, through the whole loss cdf, to a number in $[0, 1]$; this pdf depends on all the c.c.d. of l ones into n groups (we enumerate this c.c.d. with the index j). The j-th c.c.d., corresponding to the vector $\mathbf{p}_{l,n}(j)$, is related through a copula function to a joint probability. This joint probability corresponds to the probability to have a loss less than or equal to l percentage units distributed into the groups in the way explained by the j-th c.c.d. The mean of the joint probabilities computed for all the c.c.d. for the same number of ones, i.e. for the same l ones into n groups, represents the (mean) cumulated probability to count l ones, or better to have l events whose probabilities are the marginal cdf computed within the classes, if we assume that the different ways to distribute the events are equally probable. Anyway, we may consider an expectation corresponding to the assumption related to the affecting policy of the events.

Corollary 2.5 (Expected whole loss cdf linked to the n-dimensional copula). *The expected whole loss cdf linked to the n-dimensional first level dependence C, characterized by n homogeneous groups of dimensions m_s, $s = 1, \dots, n$, respectively, such that $\sum_{s=1}^{n} m_s = N$, is the function $F^{wr} : [0, N] \to [0, 1]$ such that*

$$F^{wr}(i) = \mathbb{E}\big(C(\mathbf{F}(\mathbf{p}_{l,n}(j))),\ j = 1, \dots, \hat{D}^c(i, n)\big), \tag{3}$$

where $\mathbb{E}$ is such kind of mean corresponding to the average if the distributions are equally probable, $\mathbf{p}_{l,n}(j)$ is an n-vector counting the ones corresponding to the j-th c.c.d. of l ones into n groups of dimensions m_s, $s = 1, \dots, n$, where $\mathbf{F}(\mathbf{p}_{l,n}(j))$ is the vector of margins corresponding to the j-th c.c.d. of l ones into n groups, and where $\hat{D}^c(i, n)$ counts the c.c.d. of i ones into n groups.

[8] If we consider a greater threshold, we will obviously have a greater correction for the granularity error.

[9] Here the compatibility is with respect to the cardinalities of the groups. These c.c.d. are ordered with respect to some specified criteria.

We can observe that the suggested hybrid model is a non-exchangeable model since the classes are represented by different default probabilities p^j, $j = 1, \dots, n$. This property can be regarded as a correction of the exchangeable one-factor models (see [10, 14, 15]).

Remark 2.6. We see that in order to define the right weights in (3) it is necessary to define how the defaults will distribute into the groups. If we assume that the groups are equally affected by the defaults' occurrence or that the probability that a default attaches a group i is the same as for group j (for all i, j), then the expectation operator in (3) becomes the arithmetic mean.

3 Some experiments

We consider here a loan portfolio clusterized in a finite number of classes for which we assume that the event of default occurs if the losses are greater than a fixed level. We evaluate the marginal probabilities of default of each loan and we are interested in the cdf of the percentage of default occurring that we call *whole loss distribution*.

In this example, we consider a portfolio of dimension $N = 60$ clusterized[10] into $n = 3$ classes. These groups are assumed homogeneous with marginals equal to the centroids of the same group, which represents the vector of the default probabilities:

$$\text{centroids} = \begin{bmatrix} 0.03 & 0.05 & 0.07 \end{bmatrix}.$$

The dependence parameters within the groups correspond to the levels of Kendall $\tau = 0.6, 0.43, 0.8$, while the parameter between the groups corresponds to a Kendall $\tau = 0.048$. We assume a Clayton dependence structure between the groups, meaning that the dependence parameter is $\alpha_{\text{between}} = 0.1$.

We compare the *whole loss cdf* computed in four scenarios characterized by different levels of concentration. The scenarios considered are the following:

1. *equally distributed* case, corresponding to three groups each one composed by 20 loans;
2. *not perfectly equally distributed* case, corresponding to a cardinality vector $n^2 = (10, 25, 25)$;
3. *concentration on the riskiest class* case, corresponding to a cardinality vector $n^3 = (5, 5, 50)$;
4. *concentration on the safest* class, corresponding to a cardinality vector $n^4 = (50, 5, 5)$.

We compare the mean *whole loss cdf* computed following the HHC approach for Gaussian and Clayton limiting (Gaussian/Clayton HHC for short) with that computed with a pure limiting model (PL for short) corresponding to the case of a large number of *large* classes[11]. We assume that all *small* classes are evaluated with a U-CHC model. We then properly consider U-HHC loss cdfs.

In Figures 1 and 2 we compare the *whole loss cdf* for the HHC Clayton and Gaussian model, respectively, and the corresponding PL model with the same Kendall τ into the highest level of dependence, considering different scenarios. We observe that the Clayton PL model compared to the HHC model approximates better the mean *whole loss cdf* in the *equally distributed* case and that the goodness of approximation decreases with the increase in concentration toward the riskiest class. This error is composed by a *clusterization error*, including the *exchangeable error*, which corresponds to the *sector aggregation risk*, and the *equally affecting policy error* and by a *partial error of granularity*[12]. This last kind of error is avoided in the HHC model since here we consider a not perfectly fine-grained assumption or we recover the fine grain into the *small* classes. The impact of the errors is greater for lower level of percentage losses and generally greater for the Gaussian PL model than for the Clayton one. On the other hand, the *clustering error* depends on the fact that the PL

[10] We assume to group with a clustering method, based on the Euclidean metric with a little tolerance, like $\delta = 0.09$.

[11] We observe that the first scenario, the mean *whole loss cdf* computed with the HHC approach, coincides with the LCC mean *whole loss cdf*, since in this case all classes are *large*.

[12] We talk about a *partial error of granularity* since we compare the PL with the HHC whole loss cdf that corrects the granularity only for *small* classes. The properly *error of granularity* can be recovered by a comparison between the PL and the CHC whole loss cdf.

models do not specialize the kind of distribution considered, since the loans are assumed exchangeable and equally affected, i.e. are not clustered into the groups. Moreover, we observe that the main error of the Clayton PL is due to the granularity problem since the bigger divergence seems to be with respect to the concentrated cases. This means that Clayton PL is riskier because of a big *partial error of granularity*[13] which may be corrected with an HHC model. On the contrary, the Gaussian PL approximates better the mean whole loss cdf in the concentrated cases showing that the main error is due to the clusterization aspect[14] that generates a too conservative loss cdf; HHC corrects the error with a less conservative loss cdf. Indeed, the comparison between the loss cdf in case 2 and the aggregated cases can be considered explicative of the impact of the choice of a greater number of clusters; a lower error of granularity implies a more conservative whole loss cdf.

Figures 3 and 4 explain the same comparison as before but for a different level of Kendall τ. We observe that the error of PL cdf is lower for a higher level of Kendall τ, and that the PL Clayton approximates better the *equally distributed* case for a lower level of dependence, while the concentrated case, particularly toward the riskiest class, for a higher level of dependence. We can say that Gaussian PL is less conservative if τ increases, meaning that the *clusterization error* is dominated by the *partial error of granularity* and then implying a minor correction impact of the HHC. Moreover, we observe that if the dependence between classes increases, the impact of the *partial granularity error* becomes the main component also in the Clayton case, implying a reverse order of the curves in Figure 3. Here Clayton PL is the most conservative approach; a lower clusterization error means a minor distance from the concentrated cases but a major distance with respect to case 1, that has no correction for granularity error at all.

Figures 5 and 6 compare the Clayton HHC and the Gaussian HHC mean *whole loss cdf* with the corresponding cdf computed through the PL model for a particular concentration scenario and for several dependence levels. We observe that if the correlation between classes increases, the HHC loss cdf goes toward the PL loss cdf where we have a concentration on the same class. The *clusterization error* of PL models becomes lower for higher dependence between the clusters implying that also for Clayton PL the granularity error is preponderant. The effect of a decreasing *clusterization error* is more evident if τ increases; PL loss cdf is in general more conservative than the HHC one, that however becomes more conservative if τ increases.[15]

Figures 7 and 8 compare the Clayton/Vasicek HHC mean *whole loss cdf* with the Clayton/Vasicek LCC mean *whole loss cdf* and the corresponding PL loss cdf for different concentration scenarios. We observe that in the analyzed scenarios, the LCC cdf suffers from a big *partial error of granularity* that is comparable to the approximation of the PL cdf. This implies that in this case the *clusterization error* is quite non-influential. The scenarios analyzed give the same information and the evidences are invariant for the level of dependence considered. In this scenario, that is the worst one, the HHC approach is more conservative than the others, while in the best scenario HHC produces a riskier *whole loss* cdf correcting for granularity and concentration.

In Figure 9, we plot the mean *whole loss cdf* for Clayton HHC, Clayton LCC and PL Clayton models for the *not perfectly equally distributed* case and for several levels of Kendall τ. We observe that the PL cdf explains a *clusterization error* that is dominant with respect to the *partial error of granularity*[16]. Moreover, the *partial error of granularity* decreases and the HHC loss cdf becomes riskier if the dependence level between groups increases[17].

Figures 10 and 11 compare Clayton and Gaussian HHC and Clayton and Gaussian LCC mean *whole loss cdf*, respectively, in several concentration cases. The evidence shows that the *partial error of granularity* increases with the concentration, independently of the direction of this concentration[18], since the number of

[13] This is the dominant error since the Clayton PL diverges mainly with respect to the concentrated cases that correct in a better way the granularity.

[14] In fact, Gaussian PL is nearer to those of concentrated cases where the granularity error is minor.

[15] This effect is due to a decreasing clusterization error.

[16] This effect is justified by the fact that the granularity is corrected in one class and then it has not a dominant effect.

[17] This implies that the *clusterization error* increases while the *partial error of granularity* decreases.

[18] This means that the evidence explained here is true if we consider a concentration toward the riskiest class but also toward the safest one.

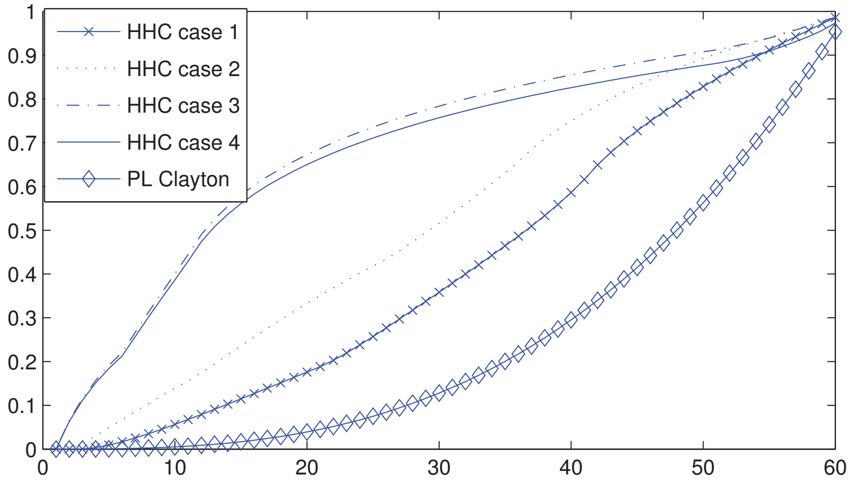

Figure 1. Clayton HHC mean *whole loss cdf* for several concentration scenarios vs PL Clayton loss cdf. Kendall $\tau = 0.048$ for PL and the highest level of Clayton HHC model.

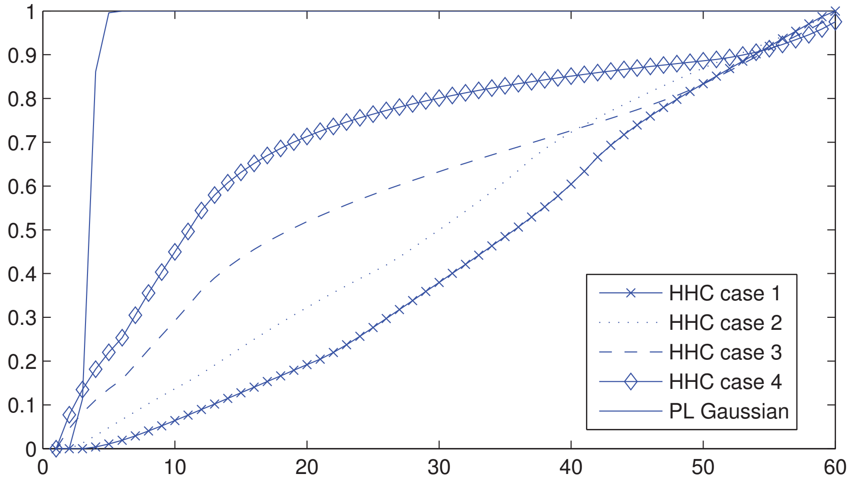

Figure 2. Gaussian HHC mean *whole loss cdf* for several concentration scenarios vs PL Gaussian loss cdf. Kendall $\tau = 0.048$ for PL and the highest level of Gaussian HHC model.

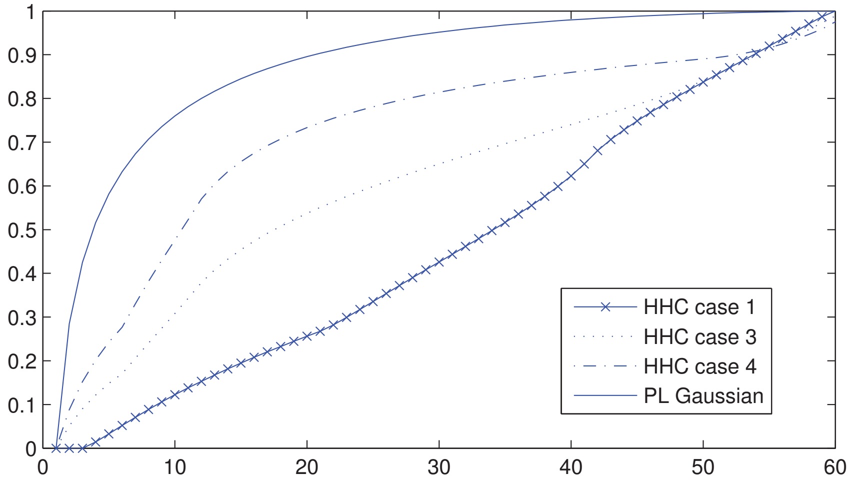

Figure 3. Gaussian HHC mean *whole loss cdf* for several concentration scenarios vs PL Clayton loss cdf. Kendall $\tau = 0.43$ for PL and the highest level of HHC model.

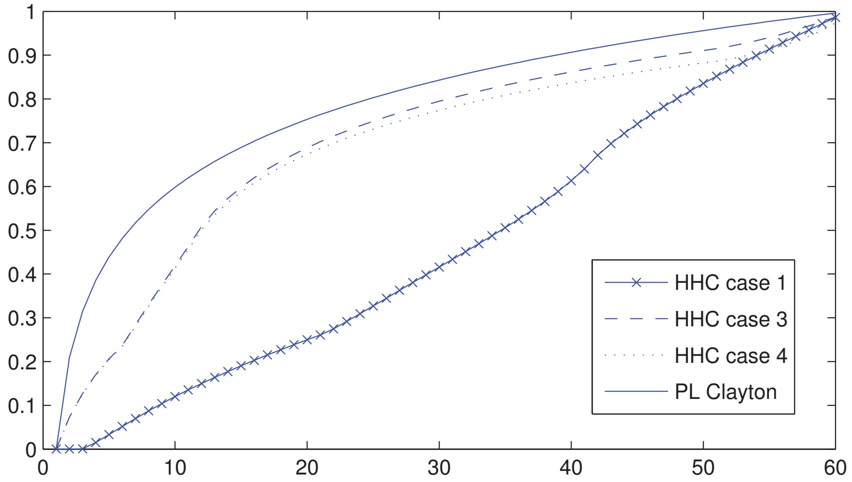

Figure 4. Clayton HHC mean *whole loss cdf* for several concentration scenarios vs PL Clayton loss cdf. Kendall $\tau = 0.43$ for PL and the highest level of HHC model.

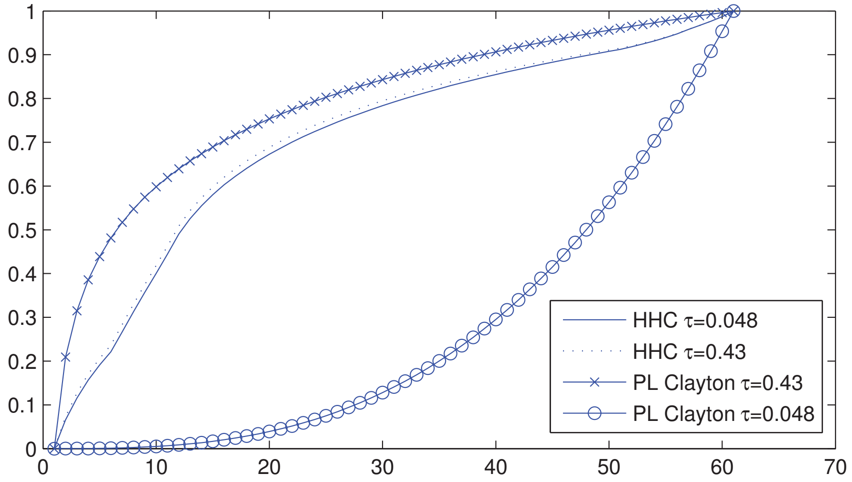

Figure 5. Clayton HHC mean *whole loss cdf* vs PL Clayton loss cdf in the concentration on the riskiest class case. Several levels of Kendall for PL and the highest level of HHC model.

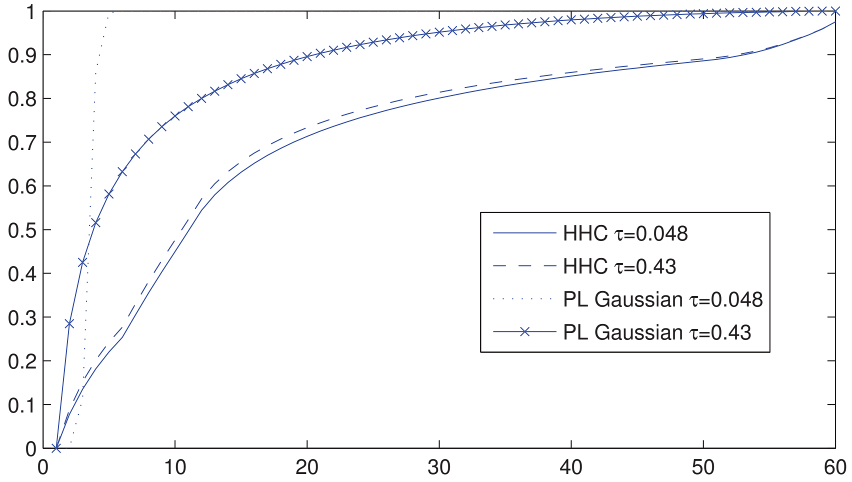

Figure 6. Gaussian HHC mean *whole loss cdf* vs PL Gaussian loss cdf in the concentration on the riskiest class case. Several levels of Kendall for PL and the highest level of HHC model.

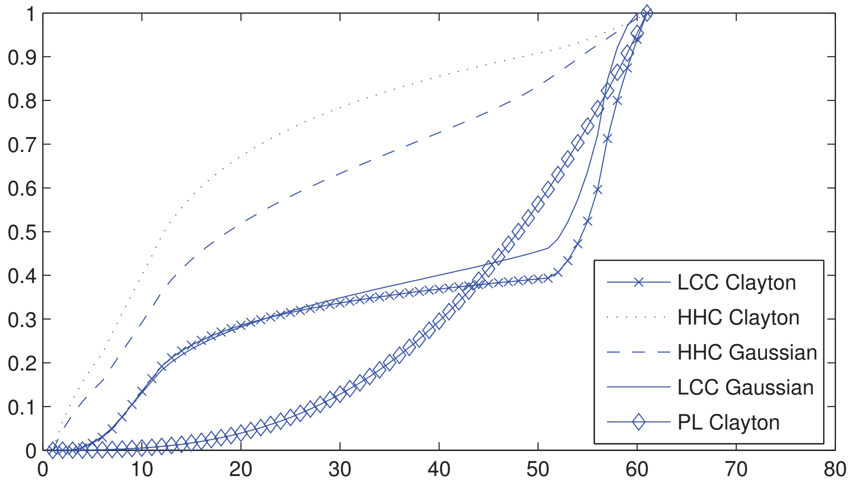

Figure 7. Clayton/Gaussian HHC vs Clayton/Gaussian LCC mean *whole loss cdf* and PL Clayton loss cdf in the concentration on the riskiest class case. Kendall $\tau = 0.048$ for PL and the highest level of HHC model.

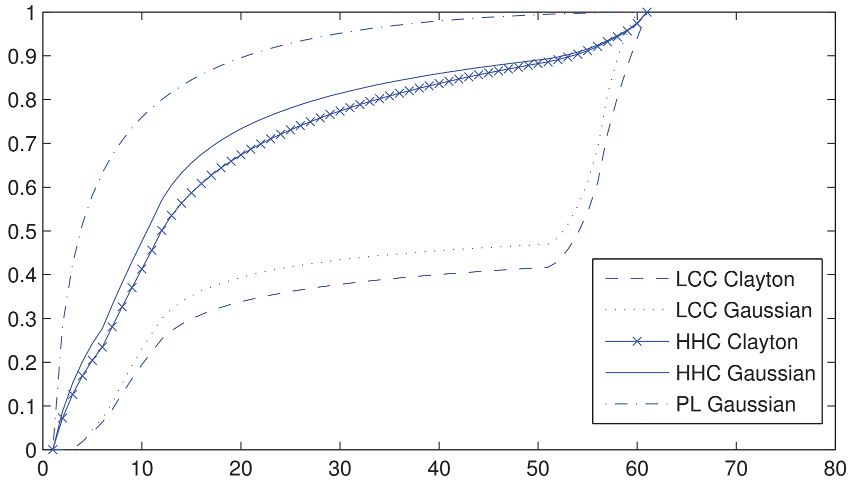

Figure 8. Clayton/Gaussian HHC vs Clayton/Gaussian LCC mean *whole loss cdf* and PL Gaussian loss cdf in the concentration on the safest class case. Kendall $\tau = 0.043$ for PL and the highest level of HHC model.

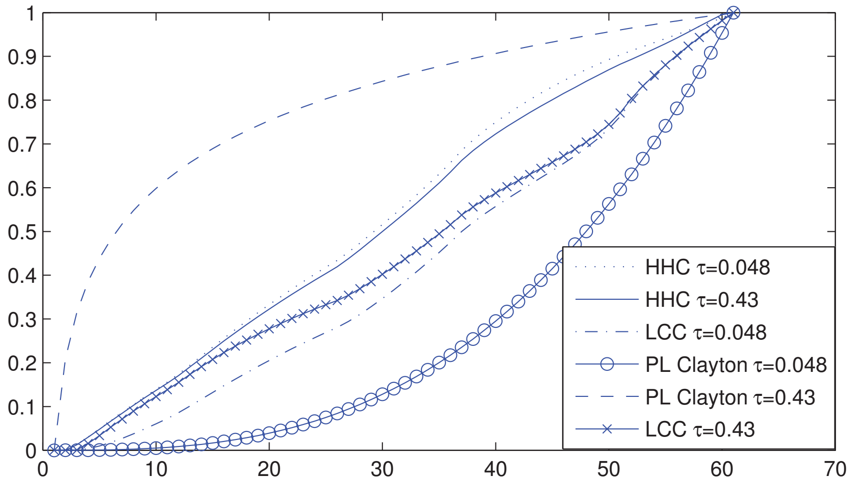

Figure 9. Clayton HHC vs Clayton LCC mean *whole loss cdf* and PL Clayton loss cdf in the *not perfectly equally distributed* case. Several levels of Kendall τ for PL and the highest level of HHC model.

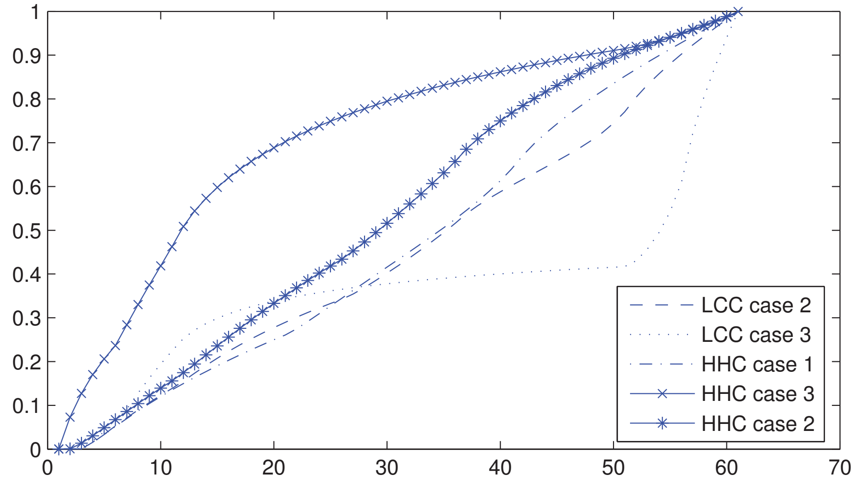

Figure 10. Clayton HHC vs Clayton LCC mean *whole loss cdf* in several concentration scenarios. Kendall $\tau = 0.43$ for the highest level of HHC and LCC models.

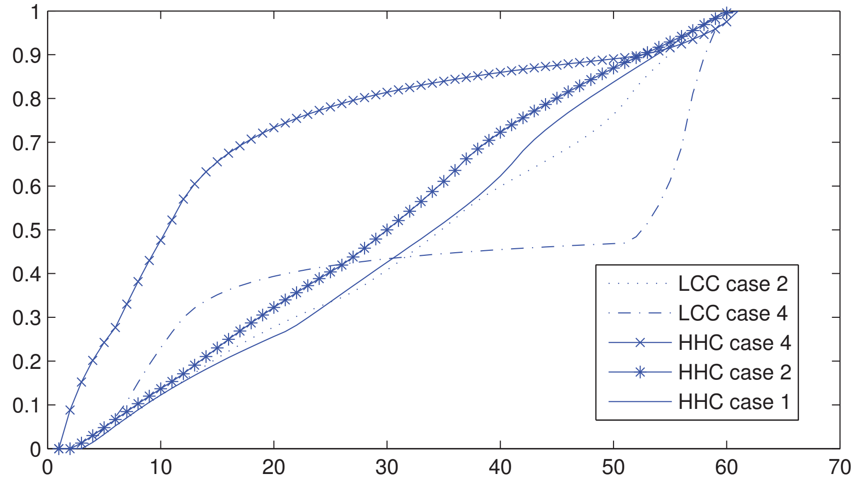

Figure 11. Gaussian HHC vs Gaussian LCC mean *whole loss cdf* in several concentration scenarios. Kendall $\tau = 0.43$ for the highest level of HHC and LCC models.

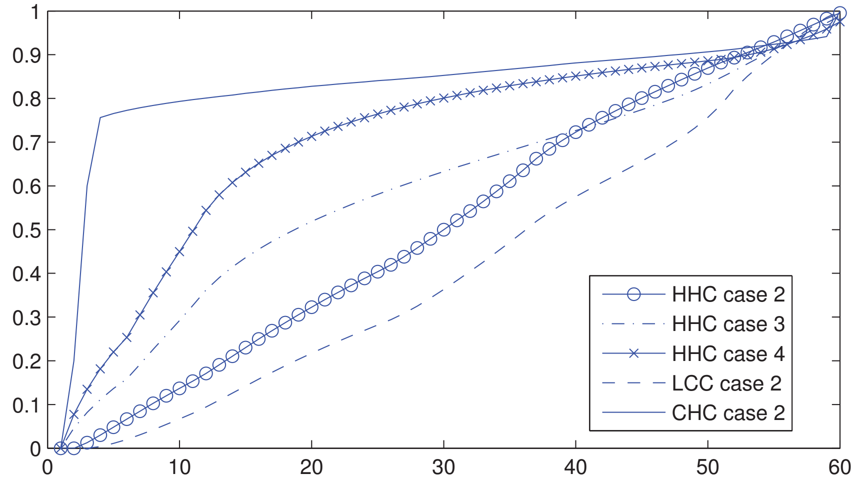

Figure 12. Gaussian HHC and Gaussian LCC mean *whole loss cdf* vs CHC in the *not perfectly equally distributed* case. Kendall $\tau = 0.048$ for the highest level of HHC, LCC and CHC models.

N. groups/conc.	CPU time	Ts Gaussian/Clayton LCC	Ts PL Gaussian/Clayton
2/eq.dist.	0.3	0.027/0.03	0.19/0.22
3/eq.dist.	1.68	0.14/1.05	1.57/1.6
3/np eq.dist.	67.89	66.35/67.26	67.78/67.81
3/conc.	34.8	33.26/34.17	34.69/34.72

Table 1. CPU computation time in seconds and CPU timesaver with respect to the Clayton HHC algorithm of the left queue (till 10 percent) of the mean *whole loss distribution* with LCC and PL approaches. The dimension of the set is $n = 60$.

small classes increases. Both figures show that the more conservative curve is obviously the HHC in the riskiest scenario for both the Gaussian and the Clayton case.

In Figure 12, we compare the Gaussian HHC and the Gaussian LCC mean *whole loss cdf* with the clusterized homogeneous Clayton copula-based (CHC for short) loss cdf (see [3]) that we consider as our benchmark, in the *not perfectly equally distributed* scenario. The difference between the LCC mean *whole loss cdf* and the CHC loss cdf represents the properly *error of granularity*, since in the LCC model all classes are considered *large*, i.e. with an infinity degree of granularity. On the other hand, the difference between the HHC mean *whole loss cdf* and the CHC loss cdf explains an error of granularity attaching only the classes classified as *large*; we call this error *semi-granularity error*. The scenario considered here contains only one small class, implying that the improvement of HHC with respect to the LCC, i.e. the *partial error of granularity*, is small. We observe also that the CHC loss cdf is better approximated by the HHC cdf for the concentrated scenarios (toward the riskiest or the safest class) since in this case the HHC model corrects the granularity of two classes. Anyway, we may conclude that if our benchmark is the CHC model in the *not perfectly equally distributed* scenario, the granularity aspect seems to be prevailing on the concentration assumptions. In other words HHC in the worst case is riskier than the CHC one, due to the *semi-granularity error* affecting only large classes.

An interesting question concerns the scalability of the suggested algorithm to compute the mean *whole loss cdf* with the HHC model, i.e. the maximum dimension to which our techniques can be extended. We analyze the scalability of the HHC algorithm with respect to the number of groups and the level of concentration. The MatLab™ algorithm was lunched on a Dell Dimension DXP061 workstation, Intel(R) 2CPU, 6400@213 GHz, 2.00 GB RAM. Table 1 reports the computing time for the first 10 percent of the *whole loss cdf* corresponding to the sample of dimension $n = 60$ for different numbers of groups and concentration assumptions and the CPU timesaver of the LCC and the PL approaches with respect to the HHC method. We consider three different concentration scenarios for the case of three clusters: i) the equally distributed scenario for the cardinalities $(20, 20, 20)$ performs three *large* classes; ii) the not perfectly equally distributed case for cardinalities $(10, 25, 25)$ or with two *large* and a *small* class; iii) a concentration case for cardinalities $(5, 5, 50)$ or for two *small* and a *large* class. We observe that in the equally distributed scenario, the HHC model coincides exactly with the LCC one, since all classes are large implying that the *partial error of granularity* coincides with the *semi-granularity error* attaching the HHC procedure.

We can observe that the number of groups is a critical variable; the CPU time increases exponentially with the number of groups. Indeed, we observe that LCC and PL approaches are faster than the HHC algorithm; the CPU timesaver corresponds to the price to correct a *partial error of granularity* as it is shown in Table 2. Here we compare the relative errors of the deciles of the *whole loss cdf* computed with different methods with respect to our benchmark that is the CHC cdf representing the best approximations for high-dimensional problems (see [3]) for a Kendall $\tau = 0.43$. We notice that when the number of *small* groups increases, the performance of the HHC method is clearly better than that of other models.

Model	AE 3/eq.dist.	AE 3/np eq.dist.	AE 3/conc.
Clayton HHC	28.01%	26.41%	6.61%
Gaussian HHC	27.94%	26.52%	14.45%
Clayton LCC	28.01%	27.86%	25.85%
Gaussian LCC	27.94%	27.98%	27.25%
PL Clayton	27.7%	27.7%	27.7%
PL Clayton	32.31%	32.31%	32.31%

Table 2. Percentage of absolute errors of the loss distributions' deciles computed with different models with respect to the CHC-based whole loss distribution in the equally distributed scenario. Dimension of the set $n = 60$ and Kendall $\tau = 0.43$.

4 Conclusions

The aim of this work is to propose a new model to determine the loss distribution allowing to take into account the non-exchangeable behavior of a portfolio clustered into several sub-portfolios or classes of homogeneous loans and to recover the fine-grained assumption of the *small* classes. The suggested model is hybrid since it unifies two different approaches, the limiting (see [24, 28, 29] for the pure limiting models and [4] for the LCC model) and the CHC-based one (see [3]). In fact, the loss distribution in each class is computed by a limiting technique if the class is *large*, or through the CHC model if it is *small*. Finally, the aggregation is trained by considering the dependence structure between the classes. The linkage of the loss distributions of each classes may be seen as the highest level of aggregation into the dependency structure of a hierarchical copula function.

In the same spirit of the multi-factor models in the literature on sector concentration risk (see [22, 27]), we analyze the impact of the *clusterization* and the *properly and semi-granularity errors* in several concentration scenarios emphasizing the correction introduced with the HHC model.

References

[1] K. Aas, C. Czado, A. Frigessi and H. Bakken, Pair-copula constructions of multiple dependence, *Insurance Math. Econom.* **44** (2009), 182–198.

[2] E. Bernardi and S. Romagnoli, Computing the volume of an high-dimensional semi-unsupervised hierarchical copula, *Int. J. Comput. Math.* **88** (2011), 2591–2607.

[3] E. Bernardi and S. Romagnoli, Clustered copula-based probability distribution of a counting variable for high-dimensional problems, *J. Credit Risk* **9** (2013), 3–26.

[4] E. Bernardi and S. Romagnoli, Limiting loss distribution on a hierarchical copula-based model, *Int. Rev. Appl. Financ. Issues Econom.* **4** (2012), 126–135.

[5] U. Cherubini, F. Gobbi, S. Mulinacci and S. Romagnoli, *Dynamic Copula Methods in Finance*, John Wiley Finance Series, Wiley, Chichester, 2012.

[6] U. Cherubini, E. Luciano and W. Vecchiato, *Copula Methods in Finance*, John Wiley Finance Series, Wiley, Chichester, 2004.

[7] A. Cifuentes, I. Efrat, J. Glouck and E. Murphy, Buying and selling credit risk: A perspective on credit-linked obligations, in: *Credit Derivatives: Application for Risk Management, Investment and Portfolio Optimization*, Risk Books, London (1999), 112–123.

[8] A. Cifuentes and G. O'Connor, The binomial expansion method applied to CBO/CLO analysis, Moody's special report, 1996.

[9] B. Choroś-Tomczyk, W. K. Härdle and O. Okhrin, Valuation of collateralised debt obligations with hierarchical Archimedean copulae, *J. Empirical Finance* **24** (2013), 42–62.

[10] S. Emmer and D. Tasche, Calculating credit risk capital charges with the one-factor model, *J. Risk* **7** (2005), 85–101.

[11] D. Duffie and N. Gârleanu, Risk and valuation of collateralised debt obligations, *Financ. Anal. J.* **57** (2001), 41–59.

[12] D. Francois, V. Wertz and M. Verleysen, Non-Euclidean metrics for similarity search in noisy datasets, in: *Proc. European Symposium on Artificial Neural Networks* (ESANN'2005), 339–344.

[13] R. Frey and A. J. McNeil, Modelling dependent defaults, preprint (2001), Department of Mathematics, ETH Zürich.

[14] M. Gordy, A comparative anatomy of credit risk models, *J. Banking Finance* **24** (2000), 119–149.

[15] M. Gordy, A risk-factor model foundation for ratings-based bank capital rules, *J. Financ. Intermed.* **12** (2003), 199–232.

[16] M. Gordy, Granularity adjustment in portfolio credit risk measurement, in: *Risk Measures for the 21st Century*, Wiley, New York (2004), 109–121.

[17] M. Hofert and M. Scherer, CDO pricing with nested Archimedean copulas, *Quantitative Finance* **11** (2011), 775–787.

[18] H. Joe, *Multivariate Models and Dependence Concepts*, Chapman & Hall, London, 1997.

[19] D. X. Li, On default correlation: A copula function approach, preprint (2000), paper 99–07, Risk Metrics Group.

[20] A. W. Marshall and I. Olkin, Families of multivariate distributions, *J. Amer. Statist. Assoc.* **83** (1988), 834–841.

[21] R. Nelsen, *Introduction to Copulas*, Springer, Heidelberg, 2006.

[22] M. Pykhtin, Multi-factor adjustment, *Risk Mag.* **17** (2004), 85–90.

[23] C. Savu and M. Trede, Hierarchical Archimedean copulas, *Quantitative Finance* **10** (2010), 295–304.

[24] P. J. Schönbucher, Taken to the limit: Simple and not-so-simple loan loss distributions, *Wilmott Mag.* **1** (2004), 63–72.

[25] P. J. Schönbucher and D. Schubert, Copula dependent default risk in intensity models, preprint (2001), Department of Statistics, Bonn University.

[26] C. Schorin and S. Weinrich, Collateralized debt obligation handbook, preprint (1998), Morgan Stanley Dean Witter.

[27] D. Tasche, Measuring sectoral diversification in an asymptotic multi-factor framework, *J. Credit Risk* **2** (2006), 33–55.

[28] O. Vasicek, Probability of loss on loan portfolio, preprint (1987), KMV Corporation.

[29] O. Vasicek, The loan loss distribution, preprint (1997), KMV Corporation.

[30] L. Wang, L. Bo and L. Jiao, A modified K-means clustering with a density-sensitive distance metric, in: *Rough Sets and Knowledge Technology*, Lecture Notes in Comput. Sci. 4062, Springer, Berlin (2006), 544–551.