No. 2010–103

MODELING WITHIN- AND ACROSS-CUSTOMER ASSOCIATION IN LIFETIME VALUE WITH COPULAS

By Nicolas Glady, Aurélie Lemmens and Christophe Croux

October 2010

ISSN 0924-7815

TILBURG UNIVERSITY

Modeling Within- and Across-Customer Association in Lifetime Value with Copulas

Nicolas Glady
ESSEC Business School

Aurélie Lemmens
Erasmus University Rotterdam

Christophe Croux
K.U.Leuven and Tilburg University

Abstract: Recent advances in linking Recency-Frequency-Monetary value (RFM) data to Customer Lifetime Value (CLV) in non-contractual settings rely on the assumption of independence between the transaction and spend processes. We propose to model jointly the inter- and intra-customer dependency between both processes using copulas, hereby accounting for the double correlation *within* and *across* customers. Applied to a unique data set of securities' transactions, we find that modeling both associations enhances the accuracy of CLV predictions, thus improving customer valuation and selection tasks.

Keywords: Association, Copula, Customer Lifetime Value, Across and Within Customers.

JEL code: M31, C51, C53.

INTRODUCTION

As firms have to comply with stricter marketing budget constraints, customer lifetime value (CLV) has become a popular metric in marketing research and practice for customer valuation, customer selection and the allocation of marketing resources over the customer base (Berger et al. 2002, Gupta et al. 2004, Kumar and Venkatesan 2006, Rust et al. 2004, Venkatesan and Kumar 2004). An accurate estimation of the future cash flows each customer is likely to generate can efficiently drive firms' decisions on how to prioritize marketing efforts on customers that are expected to provide the highest revenues (Venkatesan et al. 2007).

While several models have been proposed to compute CLV in a continuous-time non-contractual setting (see Gupta et al. 2006, for a review), probably the most prominent and successful existing approach so far is the stochastic framework of buyer behavior proposed by Fader et al. (2005b). The approach builds on the well-known Pareto/NBD framework introduced by Schmittlein et al. (1987) and links the observed recency-frequency-monetary value (RFM) measures to the unobserved customers' latent traits to predict future customer behavior. The Pareto/NBD models the flow of transactions over time in a non-contractual setting, accounting for dropout. It has been successfully applied in multiple contexts and industries (Reinartz and Kumar 2000; 2003, Schmittlein and Peterson 1994).

In order to characterize a customer in terms of his/her future cash flows, Fader et al. (2005b) specify a separate Gamma/Gamma sub-model for the amount spent per transaction. An assumption made in their CLV calculation is the independence between the transaction stream and the spend process. This assumption implies that the model of buyer behavior can be decomposed into two sub-models, and the expected CLV obtained by multiplying the number of discounted expected transactions by the expected net cash flow per transaction. They warn that this assumption might be invalidated in some applications, and call for the development of a model that would relax the independence assumption.

In this paper, we build up their framework by using *copulas* to model the dependency between the transaction flow and spend process. A copula describes the joint behavior of a multivariate random variable after controlling for the marginal behavior

of the single random variables. Recently, Danaher and Smith (2009) advocated the use of copula in marketing to model dependency structures. They prove very useful in contexts where the marginal distributions are known and association between the marginal random variables is anticipated (Danaher and Hardie 2005, Danaher and Smith 2009). Copulas allow for more flexible association structures than a bivariate distribution (e.g., log-normal distribution in Abe 2009). In particular, marginals of the multivariate distribution do not need to be of the same family. In case of independence, our copula-extended model boils down to the original CLV model proposed by Fader et al. (2005b).

We use two levels of copulas to model simultaneously the *inter-* and *intra-customer* association between the transaction and the spend processes. The former characterizes the association between the mean purchase frequency and mean transaction value *across* customers. The latter captures the association between the interpurchase time and a given transaction's value *within* each customer's transaction path. While modeling both associations is needed to avoid biased predictions, it is also managerially relevant. The inter-customer association characterizes how a *permanent* change in one of the two processes impacts the other process. For instance, marketing actions such as loyalty programs might have an enduring effect on buyers' behavior (Lewis 2004). In contrast, the intra-customer association captures the effect of a *temporary* (or one-period) change, e.g. a temporary increase in transaction value in reaction to a price promotion. By incorporating customer characteristics to model the heterogeneity in both associations, we can detect customer segments with different (inter- and intra-customer) association intensities, and thus different reactions to (different types of) marketing incentives.

We find that modeling the double association between the transaction flow and spend processes has important managerial implications. First, it enhances the accuracy of the CLV predictions, thus improving customer valuation and selection. Our results indeed show that ignoring the dependency between the transaction and the spend processes leads to an overestimation of the CLV for the customer segment that show a negative association. Second, we demonstrate that insights on the association characterizing each customer can be used for marketers' resource allocation decisions. In particular, customer segments characterized by a weak association (i.e. close to zero) turn out to be a better target for a marketing incentive than customer segments showing a strongly negative association. The reason is that, in case of a negative as-

sociation between transaction rate and transaction value, a marketing incentive that would induce an increase in transaction frequency (resp. value) would be *compensated* in part by an decrease in transaction value (resp. frequency), which would ultimately mitigate the net effect on CLV.

Lately, the question of dependency between the various components of CLV models has received an increasing attention from marketing scholars. However, to the best of our knowledge, none of them account for the double association within and across customers jointly. Borle et al. (2008) and Singh et al. (2009) offer a generalized data augmentation framework that accounts for the correlation *across* customers between both processes, while Jen et al. (2009) focuses on the temporal dependency *within* a customer between both constructs and specify a bivariate log-normal hierarchical Bayesian model of purchase timing and quantity. They find their specification to improve predictions of customers' expected income stream. Finally, focusing on the Pareto/NBD sub-model, Abe (2009) allows for a cross-sectional association between the transaction and the dropout process.

The remainder of the paper is organized as follows. In the next section, we explain and motivate the role of the intra- and inter-customer association between the transaction flow and spend process as a measure of the within- and across-customer dependency in a CLV context. Next, we introduce the concept of copulas and subsequently present the copula-extended CLV modeling framework. In the empirical application, we validate our approach on the CDNOW data used by Fader et al. (2005b), and show how the model can be used to improve CLV prediction in a new empirical application on customers' securities transactions in the retail banking sector. In turn, we then illustrate how our approach can be used to improve customer valuation, selection and resource allocation decisions. Finally, we conclude and present limitations and suggestions for future research.

INTER- AND INTRA-CUSTOMER ASSOCIATION IN CLV

At the higher level of association, the *inter-customer* association measures the relation between the mean transaction rate (frequency) and the mean transaction value

(monetary value) *across* customers. When this cross-sectional association is negative, it indicates that frequent buyers tend to spend less per transaction than infrequent buyers. That is, customers with a mean transaction rate higher than the population mean tend to spend less per transaction than customers with a mean transaction rate lower than the population mean. In contrast, a positive association would suggest that infrequent buyers also tend to spend more per transaction than infrequent buyers. While a positive inter-customer association is rather unlikely, we expect a negative inter-customer association in most cases. For instance, in the context of grocery shopping at supermarkets, a share of customers commonly visits the store once a week and makes purchases for the whole week, while another part tends to prefer daily shopping (see e.g. Bell and Lattin 1998). The existence of such customer segments suggests a negative inter-customer association. In general, the modeling of the inter-customer association is motivated by the existence of customers with different socio-demographic profiles (e.g. different lifestyles and time constraints), different price and promotion sensitivities (Ainslie and Rossi 1998) or different purchase motives (e.g. professional purpose or personal use). In a contractual setting, Borle et al. (2008) studies the case of a membership-based direct marketing company and find a substantive inter-customer association.

At the lower level of association, the *intra-customer* association, or the association *within* a customer, measures how the value of this customer's transactions depends on the time between consecutive transactions (interpurchase time) he/she makes. By analogy to the association at the higher level, we define a customer with a negative intra-customer association as one who tends to show a decrease (resp. increase) in transaction value when his/her purchase frequency temporarily increases (resp. decreases). In other words, the longer the time since this customer's last purchase, the higher the amount expected to be spent on his/her next transaction, and vice versa. A negative intra-customer association thus translates the degree to which the buying behavior of a given customer is *compensating* over time, i.e. the degree to which an increase in his/her interpurchase time is compensated or not by an increase in purchase amount. A detailed treatment of this particular type of temporal association has been recently provided by Jen et al. (2009).

At the intra-customer level, we expect the specific industry context as well as the product characteristics to induce the existence of compensating buying patterns. First, the industry context is likely to raise a different association intensity. For instance, in

the entertainment industry, purchases of music CDs are known to be driven by the releases of new titles that match customer preferences, as well as by the calendar of special occasions (e.g. Valentine day, Christmas, ...), which are likely to influence the temporal pattern of transactions. This can explain why, in the case of the CDNOW data, the association turned out to be weak. Other contexts where the intra-customer association might be low can be the temporal patterns of doctor visits in the health sector, or of night stays in the hotel industry. In contrast, in other industries where the supply is less driven by the calendar (e.g. grocery retailing), we expect a more pronounced negative intra-customer association. Charity giving is another context in which the intra-customer association can be negative and significant. For instance, van Diepen et al. (2009) find that the recency of a donation decreases the amount that is donated to the charity, which suggests a negative intra-customer association.

In addition, product characteristics are also likely to affect the intra-customer association. First, the possibility to stockpile is likely to strengthen the negative association, as consumers might decide to temporarily advance or postpone their purchase in response to available price promotions (Meyer and Assuncao 1990). In line with this argument, the degree of perishability of goods will strengthen the negative intra-customer association (Wansink and Deshpande 1994). Also, utilitarian goods are likely to show a stronger negative association than hedonic goods as the latter are generally less responsive to stockpiling than the former and promotions of hedonic goods are more likely to lead to consumption expansion than promotions of utilitarian products, which generally lead to a longer interpurchase time (Chandon and Wansink 2002).

[INSERT FIGURE 1 ABOUT HERE]

The inter- and intra-customer association can be visualized in Figure 1, which exhibits the amount spent per transaction by five imaginary customers as a function of the interpurchase time. Transactions done by different customer are represented with different symbols. Average interpurchase times and spend per transaction of each customer are depicted in bold. In this example, we observe, at the inter-customer level, a positive correlation between the average interpurchase times and average transaction values (bold symbols). At the intra-customer level, we observe three customers showing a positive slope (the crosses, lozenges, and stars), a customer with a non-significant slope (the squares), and a customer with a negative slope (the circles). Note that a positive slope between interpurchase times and transaction values should be interpreted

as a negative association between transaction rate (frequency) and transaction value, given that a transaction rate is inversely related to an interpurchase time.

The effect of the inter- and intra-customer association on the resulting CLV can be described as follows. When one assesses the effect of a change in transaction rate or transaction value on the CLV, a negative association leads to an overestimation of the change in CLV if the model does not account for dependency, while it will underestimate it in case of a positive association. In addition, the association across customers differs from the within-customer dependency in that the former captures the long-run dependency between the transaction and the spend process, while the latter measures the short-run dependency. Marketing actions can result either in a permanent increase in transaction value (e.g. a membership to a loyalty program), or in a temporary, one-period change in purchase value, e.g. due to a price promotion). The inter-customer association measures how the average transaction frequency changes in reaction to the marketing actions that have a permanent effect, while the intra-customer association measures how the interpurchase time until the next purchase will be affected by the marketing incentives that have a temporary impact.

COPULAS

Copulas can be used to model the association between two random variables X and M with marginal distributions $F(x) = P(X \leq x)$ and $G(m) = P(M \leq m)$ and joint distribution function $H(x, m) = P(X \leq x, M \leq m)$. While models for the margins F and G are commonly known, obtaining an explicit expression for the joint distribution H is generally not straightforward, motivating the use of copulas. The Sklar's theorem (Sklar 1959) yields that, for any F and G, there always exists a copula function C such that

$$H(x, m) = C(F(x), G(m)). \tag{1}$$

The copula function C is assumed to be known up to an unknown parameter θ. In

order to be a copula, the function C has to meet the following three conditions,

$$\begin{aligned} &\text{(i) } C(F,0) = C(0,G) = 0, \\ &\text{(ii) } C(F,1) = F \text{ and } C(1,G) = G, \\ &\text{(iii) if } F_1 \leq F_2 \text{ and } G_1 \leq G_2, \text{then} \\ &\quad C(F_2,G_2) + C(F_1,G_1) - C(F_2,G_1) - C(F_1,G_2) \geq 0. \end{aligned} \tag{2}$$

If f, g and h are the probability density functions corresponding to F, G and H, the copula density function c then verifies

$$h(x,m) = c(F(x),G(m))f(x)g(m). \tag{3}$$

Various families of copulas exist. The simplest one is the independent copula, which assumes the independence between X and M, given by $C(F(x),G(m)) = F(x).G(m)$. The corresponding copula density is then equal to one. One can find a plethora of other copulas in the literature (see Nelsen 2006, for more detail). The most common include the Gaussian copula, the Gumbel copula which only allows for a positive (or negative if taking the negative sign) association between X and M, as well as the Frank copula, which does not allow for extreme association values. More detail on these various copulas can be found in Appendix A.

In other to illustrate the peculiarity of each copula, Figure 2 reports the contour plots corresponding to the joint distribution of two standard-normal random variables that have a Spearman rank correlation equals to 50%, when specifying (i) an independent copula (upper-left plot), (ii) a Gaussian copula (upper-right plot), (iii) a Gumbel copula (lower-left plot), and (iv) a Frank copula (lower-right plot). While the joint distribution corresponding to the Gaussian copula is bivariate normal, the joint distribution corresponding to the Gumbel copula is asymmetric. In turn, the joint distribution corresponding to the Frank copula yields a weaker association in the tails compared to the Gaussian copula. The most appropriate copula can be selected based on goodness-of-fit measure.

[INSERT FIGURE 2 ABOUT HERE]

One of the main advantages of copulas is that they can fit complex association structures without affecting the marginal distributions. This is particularly interesting

when the distributions of the single random variables are well-known. The usefulness has been demonstrated in several marketing applications. For instance, Meade and Islam (2003) and Sriram et al. (2009) introduce copulas to model the dependency between the time of adoption of related technologies. Another marketing issue where multivariate distributions prove useful is the modeling of household's purchase timing or incidence across related - complementary or co-incidental - product categories, such as pasta and pasta sauce (Chintagunta and Haldar 1998), laundry detergent and fabric softener (Manchanda et al. 1999), or bacon and eggs (Danaher and Hardie 2005). The latter use the Sarmanov family of distribution (Sarmanov 1966), which is a special form of copulas. The Sarmanov family of distribution has also been used by Park and Fader (2004) and Danaher (2007) to model the dependency across multiple websites' browsing patterns, and more recently by Schweidel et al. (2008) to account for the correlation between acquisition and retention times across customers. Danaher and Smith (2009) demonstrates that the Sarmanov is more limited in its ability to model even moderated-sized correlation levels than the copulas described in this section. Finally, copulas have also been used on a regular basis in other research fields, in particular in finance (Cherubini et al. 2004, Glasserman and Li 2005). We refer to Danaher and Smith (2009) for a extensive overview of copulas.

MODELING CLV USING COPULAS

In this section, we outline the model for the timing and monetary value of the transactions made by individual customers. Let $m_{i,j}$ be the monetary value of the j^{th} transaction of a customer i, and $IPT_{i,j}$ be the interpurchase time preceding his/her j^{th} transaction. The assumptions on the marginal distributions of the spend process $m_{i,j}$ and the interpurchase time process $IPT_{i,j}$ are identical as in Fader et al. (2005b). In particular, the interpurchase time of a customer follows an exponential distribution with parameter λ_i, such that the total number of purchases in a unit time interval follows a Poisson distribution with expected value λ_i. On the other hand, the dollar value of a customer's transaction follows a gamma(p, ν_i) distribution, having mean p/ν_i. We call λ_i the *transaction rate* and ν_i the *revenue rate* of customer i, and both are considered as random.[1]

We expand the original CLV model in two ways. First, we introduce two levels of copula to model both the inter- and intra-customer association. Second, we extend Fader and Hardie (2007) by incorporating time-invariant covariates to account for observed customer heterogeneity in the expected transaction and revenue rate, as well as in the inter- and intra-customer association parameters.

Modeling the Association Structure

At the higher level, we define the inter-customer association as the cross-sectional association between the variables λ_i and ν_i. We capture the association through a copula distribution with parameter θ_{inter}. A negative value of θ_{inter} indicates a negative association between the average number of transactions (i.e. the frequency) and the average transaction value across customers. In other words, a customer making more transactions than the population average is likely to make lower value transactions than the population average.

At the lower level, we define the intra-customer association as the association within customer i between $IPT_{i,j}$ and $m_{i,j}$, captured through a copula distribution with parameter θ_{intra}. We parameterize the copula such that a negative θ_{intra} implies a negative association between the number of transactions and the transactions' value, or a positive association between the interpurchase time and the transactions' value. A negative value of θ_{intra} indicates that a transaction preceded by a longer interpurchase time (compared to the customer mean) is likely to be of higher value (compared to the customer mean).

Modeling Customer Heterogeneity

In line with Fader and Hardie (2007), we adapt the CLV modeling framework by incorporating time-invariant covariates to model observed customer heterogeneity. The transaction rate λ_i and the revenue rate ν_i both follow Gamma distributions with shape r and scale α, respectively shape q and scale γ. We allow the hyper-parameters α and γ to depend on a set of covariates V_i. The covariates can include general customer socio-demographics as well as company-related customer characteristics. More specifically, and following Fader and Hardie (2007), we write

$$E[\lambda_i] = \frac{r}{\alpha_i} = \frac{r}{\exp(-\rho' V_i)} \tag{4}$$

and

$$E[\nu_i] = \frac{q}{\gamma_i} = \frac{q}{\exp(\kappa' V_i)}. \tag{5}$$

The parameter ρ captures the effect of the covariates V_i on the expected transaction rate, while the parameter κ captures the effect of the covariates V_i on the expected revenue per transaction. According to (4) and (5), a positive value of ρ, respectively κ, implies a positive effect of the covariate on the transaction rate, respectively the expected transaction value.

In addition, the association between the flow of transaction and the spend process is also likely to be heterogenous over the customer base, e.g. some customer segments might exhibit a negative association, while other segments might show no significant association. We model the heterogeneity in the association parameters θ_{inter} and θ_{intra}, by specifying them as a function of customer covariates V_i

$$\theta_{inter,i} = f(\eta'_{inter} V_i), \tag{6}$$

$$\theta_{intra,i} = f(\eta'_{intra} V_i), \tag{7}$$

where f is a link function ensuring that θ_i stays within the bounds of the copula family.[2] The parameter η_{inter} captures the effect of a covariate on the strength of the inter-customer association. Therefore, it can be used to identify the socio-demographics and company-related profile of segments of customers for which the average number of transactions is weakly associated with the expected transaction value. Likewise, the parameter η_{intra} captures the effect of a covariate on the strength of the intra-customer association. It can be used to identify customer segments for which the compensating effect between the transaction flow and spend process is the least pronounced.

Estimation of the Model Parameters

The parameters to be estimated (listed in Appendix B) are collected in the vector $\boldsymbol{\theta}$. From the observed interpurchase times $IPT_{i,j}$ and transaction values $m_{i,j}$, for $1 \leq i \leq n$, with n the sample size, and $1 \leq j \leq x_i$, with x_i the number of repeated transactions made by customer i, we estimate $\boldsymbol{\theta}$ by maximizing the log-likelihood

$$\sum_i \sum_j \log L_i(IPT_{ij}, m_{ij}|\boldsymbol{\theta}). \tag{8}$$

Let f_i and g_i be the density functions of the interpurchase times and the transaction value respectively, which are allowed to be different for each customer. The corresponding distribution functions are denoted by F_i and G_i. Using Equation (3), we can write

$$\log L_i(IPT_{ij}, m_{ij}|\boldsymbol{\theta}) = \log f_i(IPT_{ij}|\boldsymbol{\theta}) + \log g_i(m_{ij}|\boldsymbol{\theta}) + \log c_{intra}(F_i(IPT_{ij}), G_i(m_{ij})|\boldsymbol{\theta}), \tag{9}$$

with c_{intra} the density of the specified copula distribution (see Appendix A).

Following the semi-parametric maximum likelihood approach for copula estimation (see for example Genest et al. 1995), we approximate $F_i(IPT_{i,j})$ by $\hat{F}_{ij} = R_{ij}/x_i$, where R_{ij} is the rank of the jth interpurchase time, taken over all x_i observed values of IPT_{ij}, and similarly for $G_i(m_{i,j})$. As such, the third term in (9) solely depends on the intra-customer association, and the parameter η_{intra} can be estimated separately from the other ones:

$$\hat{\eta}_{intra} = \underset{\eta_{intra}}{\operatorname{argmax}} \sum_i \sum_j \log c_{intra}(\hat{F}_{ij}, \hat{G}_{ij}|\eta_{intra}). \tag{10}$$

The other hyperparameters in $\boldsymbol{\theta}$ are then estimated by maximizing

$$\int\int \sum_i \sum_j \{\log f_i(IPT_{ij}|\lambda, \boldsymbol{\theta}) + \log g_i(m_{ij}|\nu, \boldsymbol{\theta})\} h_i(\lambda, \nu|\boldsymbol{\theta}) d\lambda d\nu \tag{11}$$

with h_i the joint density of the transaction and revenue rate for customer i. This joint density h_i is the product of two Gamma densities and the inter-customer copula density c_{inter}, the latter depending only on the parameter η_{inter}.

Explicit expressions for $\tilde{f}_i(\lambda, \boldsymbol{\theta}) = \sum_j \log f_i(IPT_{ij}|\lambda, \boldsymbol{\theta})$ and $\tilde{g}_i(\nu, \boldsymbol{\theta}) = \sum_j \log g_i(m_{ij}|\nu, \boldsymbol{\theta})$ are given in Fader and Hardie (2005), and were shown to depend only on the frequency, the recency, the cohort and the average of the past transaction values for the ith customer. Expression (11) involves an integration, which cannot be solved analytically. Therefore, we use Simulated Maximum Likelihood (SML), a standard econometric estimation technique (see e.g. Green 2003, pp. 590-594). To do so, we generate random draws $(\lambda^*_{i,s}, \nu^*_{i,s})$ from the bivariate distribution $h_i(\cdot|\boldsymbol{\theta})$, for $s = 1, \ldots, S = 1000$, and approximate the integral in (11) by the corresponding Monte-Carlo average. More

specifically, the SML estimator for $\boldsymbol{\theta}$ maximizes

$$\frac{1}{S}\sum_{s=1}^{S}\sum_{i=1}^{n}\{\tilde{f}_i(\lambda_{i,s}^*,\boldsymbol{\theta})+\tilde{g}_i(\nu_{i,s}^*,\boldsymbol{\theta})\},$$

and its computation requires a numerical optimization routine.

The fact that the likelihood can be split in two parts allows to study the two levels of association separately. If data at the individual transaction level of information are not available (as it is the case in the first empirical application below), only the inter-customer association can be estimated. Finally, the estimation procedure also accounts for the attrition or "death" process (see Appendix B).

CLV Prediction

The customers' value at horizon H for customer i, $CLV_{i,H}$, is given by the discounted sum of all net revenues that will be generated within the next H time units. Let T denote the time the prediction is made, then future transactions may take place at time points $T+t_1$, $T+t_2$, ... with corresponding transaction values $m_{t_1}, m_{t_2}, \ldots$. As in Gupta et al. (2004), we define

$$CLV_{i,H}=\sum_{t_j\leq H}\frac{margin\times m_{t_j}}{(1+d)^{t_j}}, \tag{12}$$

where $margin$ is a gross margin, supposed constant, and d stands for the discount rate. Note that the $CLV_{i,H}$ defined in (12) is a random variable.

Once the model is estimated, it is possible to simulate future transaction streams for every customer, resulting in the simulated distribution of the CLV over the next H periods. Our approach is similar in spirit to Singh et al. (2009). Details are provided in Appendix C. Finally, the average over the simulated distribution yields a prediction of the expected CLV for this customer. Since we simulate the whole distribution of $CLV_{i,H}$, it is also possible to construct prediction intervals.

EMPIRICAL APPLICATIONS

We apply the copula-extended CLV modeling approach to two different datasets, the CDNOW data[3] and a unique dataset containing securities transactions data provided by an anonymous well-established international financial service institution. The implementation of the Pareto/NBD sub-model is based on Fader et al. (2005a).

In order to assess which family of copula is most appropriate for each application, we estimate the copula-extended model using the independent, Gauss, Gumbel and Frank copulas and compare their fit and out-of-sample predictive performance.

CLV in E-Commerce: the Case of CDNOW

The CDNOW data contains transactions of customers on the online music site CDNOW (Fader et al. 2005b). The transaction data (number of repeated transactions, recency, average transaction value and cohort) cover 78 weeks for a sample of 2,357 CDNOW customers who made their first-ever purchase at the website during the first quarter of 1997. We use the first 39 weeks of 1997 for the estimation of the model parameters, and keep the remaining 39 weeks as hold-out sample to assess the predictive performance of the model. We apply the same margin as the original paper, i.e. 30%. Note that we do not include an intra-customer association copula, nor do we model the heterogeneity using additional covariates as the data do not contain additional covariates. We thus have θ_{inter} equal across all customers.

We find that the Gauss copula yields the highest log-likelihood (LL) ($LL = 53,815.55$ for the independent; $LL = 56,088.19$ for the Gauss; $LL = 53,873.46$ for the Gumbel; $LL = 54,631.45$ for the Frank). In particular, it provides a substantial improvement compared to the independent model, suggesting that accounting for a non-zero inter-customer association improves the model fit. Also, the out-of-sample predictive performance of the copulas allowing for dependent transaction and spend processes all outperform the independent version as the lower root mean squared errors (RMSE) testify ($RMSE = 23.80$ for the independent; $RMSE = 22.15$ for the Gauss; $RMSE = 21.66$ for the Gumbel; $RMSE = 21.69$ for the Frank).

Table 1 reports the parameter estimates of the copula-extended model for the various types of copulas, together with their significance level.[4] Under the Gauss association model, the estimated inter-customer association parameter η_{inter} equals to .41, which corresponds to $\theta_{inter} = 0.25$ (see transformation in footnote 2) and a Spearman's

correlation of 23.71%. The Frank copula yields a very similar Spearman's correlation of 23.05%. This moderately positive association is in line with the Pearson's correlation between the average transaction value and the number of transactions of 11.39% reported in Fader et al. (2005b).[5]

[INSERT TABLE 1 ABOUT HERE]

CLV in Brokerage: the Case of Securities Transactions Data

Our second application is based on the securities transactions data provided by a major international financial service institution. The data contain securities transactions made by 2,500 randomly-selected customers who made their first transaction between January 2001 and December 2003. Transactions include the purchase and selling of stocks, bonds, mutual funds, derivatives, and similar products between January 2001 and December 2005.[6] We keep the last two years (January 2004 to December 2005, $H = 24$ months) as hold-out sample to assess the predictive performance of the CLV models.

Following a common rule of thumb in business practice, we compute the monetary value of a transaction as 1% of the average amount exchanged at each transaction. In addition, we take as discount rate the weighted average cost of capital disclosed in the 2004 financial statement of the financial service provider, that is $d = 8.92\%$ on a year basis, or a monthly discount rate of 0.71%.

The data also contain socio-demographics and company-related customer characteristics used to model the heterogeneity between customers in the parameters. Descriptive statistics for these covariates are reported in Table 2. Customer characteristics include the age of the customer (which will be mean-centered in the model estimation), as well as a dummy variable accounting for the type of area where a customer is living. This variable takes the value one when the customer lives in the suburb of a city, and zero otherwise. As company-related customer covariates, we include the cohort a customer belongs to (which will also be mean-centered in the model estimation). In addition, we also include a dummy variable taking the value one when the bank is the customer's primary bank.

[INSERT TABLE 2 ABOUT HERE]

We estimate the copula-extended CLV model using the various types of copulas described earlier and find that the Gumbel copula yields the best model fit (i.e. $LL = -15,657.37$ for the independent; $LL = -15,047.05$ for the Gauss; $LL = -14,696.28$ for the Gumbel; $LL = -15,436.50$ for the Frank) for the securities transaction data. In addition, the Gumbel copula also yields a slightly superior out-of-sample predictive performance than the other specifications ($RMSE = 313.10$ for the independent; $RMSE = 314.81$ for the Gauss; $RMSE = 309.94$ for the Gumbel; $RMSE = 314.06$ for the Frank).

According to the Gumbel copula, the average inter-customer Spearman's association between the transaction flow and the spend process amounts to -49.45% (and a median value of -37.68%). This highly negative association informs us that the frequent buyers of the financial institution under study tend to spend substantially less per transaction than the sporadic buyers. In particular, we find that the top 10% most frequent buyers spend on average 2,138.98 Euros on a transaction while the 10% least frequent buyers spend on average 2,569.90 Euros per transaction. In other words, a customer spending little compared to others is also likely to make more frequent transactions than others, and vice versa.

Not accounting for the dependency structure between the spend and transaction processes leads to an overestimation of the CLV. We indeed find that the individual CLV predicted by the independent model are higher than those given by the Gumbel model for 94.51 % of the customers with a Spearman correlation of at least 10 % (in absolute value). For customers with an intra-customer correlation below 10 %, we find almost no difference between the CLV estimates of both models. This result confirms that, when accounting for the dependency between the transaction and spend processes, the resulting CLV decreases in presence of a negative association. It also highlights the relevance of the copula-extended model for customer valuation tasks.

At the lower level of association, the Gumbel copula model yields a negative average intra-customer Spearman's association of (-)3.73%. This result is in line with the recent findings of Jen et al. (2009) who found average temporal association within customers between 0% and 10% in their applications. This indicates that clients of the banking service provider do not exhibit a strong compensating buying behavior. Information on when (resp. how much) they buy (at a given transaction) is not a strong predictor of how much (resp. when) they actually buy. This is a valuable insight from a marketing perspective as offering those customers an opportunity to buy earlier than they would

usually plan to does not imply that they will spend less. The industry context is likely to drive this relatively small degree of association. Customers might choose to buy or sell depending on the fluctuations on the stock market, rather than following a personal agenda. Note that the correlation between the estimated inter- and intra-customer association parameters amounts to 18.95%, highlighting that both copulas complement each other in accounting for a different kind of association.

In Table 3, we report the parameter estimates, with their respective significance levels, for the best-fit (Gumbel) model.[7] In the first panel, we report the main model parameters, which all turn to be significant. Next to it, we report the effect of the customer covariates for the Pareto/NBD and Gamma/Gamma sub-models (see Equations 4 and 5).

[INSERT TABLE 3 ABOUT HERE]

We find that the rate at which customers make transaction decreases with the duration of their relationship with the bank (cohort), and is lower for customers living in suburbs than elsewhere. In turn, this rate is higher for customers for which the bank is their primary financial service provider. In addition, the average value of transactions customers make tends to be higher for primary-bank clients and for clients living in the suburb of a city than elsewhere.

Finally, the last panel of Table 3 informs us about the determinants of the inter- and intra-customer association. This information can be used to determine which customer segments are characterized by a strong vs. weak association. Given the intercepts' value η_{inter} and η_{intra}, an hypothetical customer having all covariates equal to zero (i.e. mean age and mean cohort as these covariates have been mean-centered, not living in suburbs, non-primary bank) would have $\theta_{inter} = 1.00$ and $\theta_{intra} = 1.03$ (see transformation in footnote 2), which correspond to an inter- and intra-customer Spearman's correlations of -0.19% and -4.50%.

Departing from this customer segment, we can assess which customer characteristics affect the inter-customer association. The positive coefficient of age ($\eta_{inter} = 1.72$) indicates that the older customers the stronger the negative relationship between purchase frequency and transaction value. To interpret this effect, let us consider two customer segments: a segment of relatively old customers and a segment of relatively young customers. In the first segment, we expect to find frequent customers to spend far less per transaction than infrequent customers, while the difference between frequent and infrequent customers will be less prominent in the second segment. Among

the youngest customers, frequent and infrequent customers will spend about the same per transaction. We are thus more likely to overestimate the CLV of older customers than younger ones. Furthermore, younger customers will also show, on average, larger changes in CLV than older ones for a similar change either in their transaction rate, or in their transaction value. This makes this group a potentially interesting target for marketing incentive. For instance, a 3 unit increase in the expected number of transactions amongst the 50% youngest customers yields an average estimated CLV increase of 32.85 Euros (corresponding to a relative change of 341.82%), while the same increase amongst the 50% oldest customers leads to an average increase of 26.98 Euros only (that is, 123.43% relative change). This illustrates that information on the association structure can be fruitful in marketing resources allocation decisions (as further illustrated in the next section).

In addition to age, customers living in suburbs of cities ($\eta_{inter} = .43$) and secondary customers ($\eta_{inter} = -1.70$) also exhibit stronger negative inter-customer association than the others. Likewise, customers belonging to a younger cohort show a stronger negative association ($\eta_{inter} = -.38$) than customers acquired a longer time ago. In conclusion, these results make the young, primary customers not living in suburbs and belonging to older cohorts potentially attractive targets for marketing incentives as an induced increase of their spending should have a larger impact on their CLV than it is the case with others.

[INSERT FIGURE 3 ABOUT HERE]

Turning to the intra-customer association, we can assess whether some customer segments are more inclined to compensating buying behavior (negative intra-customer association) than others. Younger customers ($\eta_{intra} = .04$), living in suburbs of cities ($\eta_{intra} = -.37$), issued from older cohorts ($\eta_{intra} = -.02$), for which the firm under study is their primary bank ($\eta_{intra} = -.64$) tend to exhibit the least compensating behavior. When they happen to make a transaction earlier (resp. later) than they usually do, this group of customers would not per se spend less (resp. more) on this transaction. As we mentioned before, it renders this group attractive for a marketing incentive to generate temporarily extra income for the company. These effects can be clearly visualized from the conditional histograms of the Spearman's intra-customer association. Figure 3 exhibits the conditional histogram for age (divided into four quartiles) on the left-hand side and living area (for both types of living area) on the

right-hand side. We see that the older the customers (the lighter the bars) the more the distribution of the intra-customer association moves to the left. Likewise, the distribution for customers living in suburbs (light bars) is much closer to zero than for the customers living elsewhere (dark bars).

MANAGERIAL IMPLICATIONS FOR CUSTOMER SELECTION AND RESOURCE ALLOCATION DECISIONS

The copula-extended model offers a number of benefits for managerial use over the classical CLV modeling framework proposed by Fader et al. (2005b). In this section, we illustrate these benefits further using the securities transaction data.

Customer Valuation and Selection Decisions: Improving CLV Predictions

Customer lifetime valuation is often used as a metric to assess which customers should be acquired, grown and retained (e.g. Reinartz and Kumar 2003, Rust et al. 2004). In this context, it is important for firms to obtain accurate CLV predictions. In the previous section, we found that modeling the association structure improves the accuracy of the CLV predictions and that the independent model tends to overestimate the individual CLV's when the association is negative.

Accounting for the association structure can also improve customer selection decisions. Suppose, for instance, that the financial institution under study wants to identify and select its best customers in terms of CLV. We consider different thresholds going from the 99% highest CLV percentile to the 55% highest CLV percentile. We then assess how many customers are selected by each of these selection rules (i.e. those for which the estimated CLV is higher or equal to the above-mentioned threshold) while they should not have been selected according to their *actual* CLV. Such errors can be viewed as a kind of Type II errors (Malthouse and Blattberg 2005). Table 4 reports them for both the Gumbel and independent copulas. The number of incorrectly se-

lected customers is consistently lower when using the Gumbel copula. The difference is due to the negative association between the transaction flow and spend process, which leads to an overestimation of the CLV when ignored. In conclusion, our results highlight the relevance of the copula-extended approach for customer valuation and selection tasks.

[INSERT TABLE 4 ABOUT HERE]

Resource Allocation Decisions: Assessing the Net Effect of Marketing Efforts on CLV

While copulas improve customer valuation and selection tasks, they are also beneficial for marketing resource allocation decisions. Among other authors, Berger et al. (2002) and Venkatesan and Kumar (2004) suggest to assess the impact of marketing efforts on the CLV of each customer and to allocate resources across customers (or target customers) such that the total CLV over the customer base will be maximized. Gupta and Zeithaml (2006) have shown that marketing decisions based on CLV improve firms' financial performance. In this spirit, we propose to incorporate the association between the transaction and the spend processes into the resource allocation decisions. Indeed, we argue that a negative association mitigates the net impact of a marketing incentive on CLV. If such association would be ignored, the total net effect on the CLV would be lower than expected and the resource allocation decisions might be sub-optimal.

To illustrate this point, we imagine a marketing incentive that would supposedly lead to a permanent of increase of 3 units in the expected number of transactions of customers.[8] We can then compute a CLV with and without the marketing incentive. The difference is the expected return on CLV of the action. One can then decide to target customers who show an expected CLV return at least larger than the action cost, which we assume to amount to 10 Euros per customer targeted. We then compare the resulting total CLV gain when selecting the customers using the independent model versus the Gumbel model.

Under the independent model, we select 2,151 customers out of 2,500, and the total action cost amounts to 21,510 Euros. In contrast under the Gumbel model, owing to the mitigating effect of the copula on the CLV, only 1,804 customers cross the threshold of 10 Euros, and the resulting total action cost amounts to 18,040 Euros. We then estimate the return of this hypothetical marketing action targeted at these

2,151 vs. 1,804 customers (estimated using the best-fit model). The 2,151 customers selected under the independent model yield an estimated total net return (i.e. total CLV increase minus incentive cost) of 51,392 Euros while the smaller customer subset of 1,804 customers selected under the Gumbel model yields an estimated total net return of 52,433 Euros. The Gumbel model thus offers an additional return of 1,041 Euros with a smaller scope of action (i.e. less customers have to be reached).

CONCLUSIONS

While the Pareto/NBD, Gamma/Gamma framework developed by Fader et al. (2005b) has been successfully adopted as a powerful customer valuation method for non-contractual settings, the potential association between the transaction and spend processes has led to some modeling challenges, which have been recently considered by Borle et al. (2008), Jen et al. (2009) and Singh et al. (2009). In this paper, we propose a unique approach to account for both association between the average transaction frequency and average transaction value across customers, as well as the dependency within each individual customer. We also account for the observed customer heterogeneity in both association levels through the inclusion of covariates, making it possible to identify customer segments with different degrees of association.

Using securities transaction data, we find that customers tend to show a negative association between their transaction frequency and the value of their transactions. However, it is interesting to note that, while the association across customers is large, the association at the intra-customer level turns out to be modest. We show that our copula-extended framework improves the accuracy of customer valuation tasks and, thus, positively impacts customer selection decisions. Finally, we explain that the investigation of the association structure between transaction flow and spend process help marketing resource allocation decisions, by pointing out which customers exhibit the strongest reaction to a change in one of the CLV components.

Our study suffers from a number of limitations that open areas for future research. First, our model does not account for the potential association between the dropout process and the transaction process, as done in Abe (2009). The aim of this paper was to focus on the mechanisms of association between the transaction and spend processes

because the existence of compensating buying behavior is of theoretical and managerial interest. The extension of the bivariate copula model to a trivariate copula (Danaher and Smith 2009) to account for the potential association between the transaction, dropout and spend processes is an interesting area for further development. Another limitation of our empirical application is the lack of marketing-mix variables, which would allow us to incorporate each customer's responsiveness to marketing efforts in the resource allocation decisions (Reinartz and Venkatesan 2008). An interesting extension could be to extend our approach to the framework proposed by Venkatesan and Kumar (2004) for cases when marketing-mix variables are available. Third, our model focuses on the inter- and intra-customer association structure between the transaction flow and the spend process but does not account for the possibility of a lead-lag relationship between both interpurchase times and transaction values. Such an extension would clearly offer interesting additional insights.

Acknowledgment: We thank ING Belgium for their support and useful information. This work is part of the Ph.D. thesis of the first author. All authors contributed equally to this paper.

References

Abe, M., 2009. "Counting Your Customers" One by One: A Hierarchical Bayes Extension to the Pareto/NBD Model. Marketing Science 28 (3), 541–553.

Ainslie, A., Rossi, P. E., 1998. Similarities in choice behavior across product categories. Marketing Science 17 (2), 91–106.

Bell, D. R., Lattin, J. M., 1998. Behavior and consumer preference for store price format: Why "large basket" shoppers prefer EDLP. Marketing Science 17 (1), 66–88.

Berger, P. D., Bolton, R., Bowman, D., Briggs, E., Kumar, V., Parasuraman, A., Terry, C., 2002. Marketing actions and the value of customer assets: A framework for customer asset management. Journal of Service Research 5 (1), 39–54.

Borle, S., Singh, S., Jain, D., 2008. Customer Lifetime Value Measurement. Management Science 54 (1), 100–112.

Chandon, P., Wansink, B., 2002. When are stockpiled products consumed faster? A convenience salience framework of postpurchase consumption incidence and quantity. Journal of Marketing Research 39 (3), 321–335.

Cherubini, U., Luciano, E., Vecchiato, W., 2004. Copula Methods in Finance. John Wiley Sons Inc.

Chintagunta, P. K., Haldar, S., 1998. Investigating purchase timing behavior in two related product categories. Journal of Marketing Research 35 (1), 43–53.

Danaher, P. J., 2007. Modeling page views across multiple websites with an application to internet reach and frequency prediction. Marketing Science 26 (3), 422–437.

Danaher, P. J., Hardie, B. G. S., 2005. Bacon with your eggs? Applications of a new bivariate beta-binomial distribution. The American Statistician 59 (4), 282–286.

Danaher, P. J., Smith, M. S., 2009. Modeling multivariate distributions using copulas: Applications in marketing. Marketing Science Forthcoming.

Fader, P. S., Hardie, B. G. S., 2005. A note on deriving the Pareto/NBD model and related expressions. http://brucehardie.com/notes/009/.

Fader, P. S., Hardie, B. G. S., 2007. Incorporating time-invariant covariates into the Pareto-NBD and BG/NBD models. Working Paper.

Fader, P. S., Hardie, B. G. S., Ka Lok Lee, 2005a. A note on implementing the Pareto/NBD model in matlab. http://brucehardie.com/notes/008/.

Fader, P. S., Hardie, B. G. S., Ka Lok Lee, 2005b. RFM and CLV: Using iso-value curves for customer base analysis. Journal of Marketing Research 42 (4), 415–430.

Genest, C., Ghoudi, K., Rivest, L.-P., 1995. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82 (3), 543–552.

Glasserman, P., Li, J., 2005. Importance sampling for portfolio credit risk. Management Science 51 (11), 1643–1656.

Green, W. H., 2003. Econometric Analysis, 5th Edition. Prentice Hall, New York.

Gupta, S., Hanssens, D., Hardie, B., Kahn, W., Kumar, V., Lin, N., Ravishanker, N., Sriram, S., 2006. Modeling customer lifetime value. Journal of Service Research 9 (2), 139–155.

Gupta, S., Lehmann, D. R., Stuart, J. A., 2004. Valuing customers. Journal of Marketing Research 41 (1), 7–18.

Gupta, S., Zeithaml, V., 2006. Customer metrics and their impact on financial performance. Marketing Science 25 (6), 718–739.

Jen, L., Chou, C.-H., Allenby, G. M., 2009. The importance of modeling temporal dependence of timing and quantity in direct marketing. Journal of Marketing Research 46 (4), 482–493.

Kumar, V., Venkatesan, R., 2006. Customer Relationship Management: A Databased Approach. John Wiley, New York.

Lewis, M., 2004. The influence of loyalty programs and short-term promotions on customer retention. Journal of Marketing Research 41 (August), 281–292.

Malthouse, E. C., Blattberg, R. C., 2005. Can we predict customer lifetime value? Journal of Interactive Marketing 19 (1), 2–16.

Manchanda, P., Ansari, A., Gupta, S., 1999. The shopping basket: A model for multicategory purchase incidence decisions. Marketing Science 18 (2), 95–114.

Meade, N., Islam, T., 2003. Modelling the dependence between the times to international adoption of two related technologies. Technological Forecasting and Social Change 70 (8), 759–778.

Meyer, Assuncao, 1990. The optimality of consumer stockpiling strategies. Marketing Science 9 (1), 18–40.

Nelsen, R., 2006. An Introduction to Copulas. Springer, New York.

Park, Y.-H., Fader, P. S., 2004. Modeling browsing behavior at multiple websites. Marketing Science 23 (3), 280–303.

Reinartz, W. J., Kumar, V., 2000. On the profitability of long-life customers in a non contractual setting: An empirical investigation and implications for marketing. Journal of Marketing 64 (4), 17–35.

Reinartz, W. J., Kumar, V., 2003. The impact of customer relationship characteristics on profitable lifetime duration. Journal of Marketing 67 (1), 77–99.

Reinartz, W. J., Venkatesan, R., 2008. Decision models for customer relationship management. In: Wierenga, B. (Ed.), Handbook of Marketing Decision Models. Springer Science, pp. 291–326.

Rust, R. T., Lemon, K. N., Zeithaml, V. A., 2004. Return on marketing: Using customer equity to focus marketing strategy. Journal of Marketing 68 (1), 109–127.

Sarmanov, O. V., 1966. Generalized normal correlation and two-dimensional frechet classes. Doklady (Soviet Mathematics) 168, 596–599.

Schmittlein, D. C., Morrison, D. G., Colombo, R., 1987. Counting your customers: Who are they and what will they do next? Management Science 33 (1).

Schmittlein, D. C., Peterson, R. A., 1994. Customer base analysis: An industrial purchase process application. Marketing Science 13 (1).

Schweidel, D. A., Fader, P. S., Bradlow, E. T., 2008. A bivariate timing model of acquisition and retention. Marketing Science 27 (5), 829–843.

Singh, S. S., Borle, S., Jain, D. C., 2009. A generalized framework for estimating customer lifetime value when customer lifetimes are not observed. Quantitative Marketing and Economics 7 (2), 181–205.

Sklar, A., 1959. Fonctions de répartition à n dimensions et leurs marges. Publications de l'Institut Statistiques de l'Université de Paris 8, 229–231.

Sriram, S., Chintagunta, P. K., Agarwal, M. K., 2009. Investigating consumer purchase behavior in related technology product categories. Marketing Science Forthcoming.

van Diepen, M., Donkers, B., Franses, P.-H., 2009. Dynamic and competitive effects of direct mailings: A charitable giving application. Journal of Marketing Research 46 (1), 120–133.

Venkatesan, R., Kumar, V., 2004. A customer lifetime value framework for customer selection and resource allocation strategy. Journal of Marketing 68 (4), 106–125.

Venkatesan, R., Kumar, V., Bohling, T., 2007. Optimal customer relationship management using bayesian decision theory: An application for customer selection. Journal of Marketing Research 44 (November), 579–594.

Wansink, B., Deshpande, R., 1994. Out of sight, out of mind: Pantry stockpiling and brand-usage frequency. Marketing Letters 5 (1), 91–100.

Notes

[1]A complete listing of all model assumptions is given in Appendix B.

[2]In particular, $f(\eta' V_i) = 2/\pi \arctan(\eta' V_i)$ for the Gaussian copula, $f(\eta' V_i) = exp(\eta' V_i) + 1$ for the Gumbel copula and $f(\eta' V_i) = \eta' V_i$ for the Frank copula.

[3]We thank Bruce Hardie who kindly provided us the data set.

[4]Note that the parameter estimates for the independent copula are the same as the parameters reported in Fader et al. (2005b), confirming that the specification of an independent copula function boils down to the original model specification without copula.

[5]Note that the Pearson's correlation is a poor measure of dependency when the marginals are not normally distributed (Danaher and Smith 2009), which explains the difference between our results and the Pearson's correlation.

[6]Note that we remove the automated pension plan transactions from the data set.

[7]Detailed results for the other copulas are available upon request.

[8]Note that we do not have marketing-mix data available, preventing us to assess each customer's responsiveness to marketing efforts. However, our exercise intends to show that ignoring the potential association yields different conclusions as to the effect of a change in the number of transactions or in the transaction value on the expected CLV. To that extent, the lack of actual marketing-mix information does not harm our argument. Our approach can be extended to the framework proposed by Venkatesan and Kumar (2004) for cases where marketing-mix data are available.

Table 1: Comparison of the models' parameter estimates for the CDNOW data using the independent (idpt), Gauss, Gumbel and Frank copulas. Significance levels are indicated with * for p-values lower than .1, ** for p-values lower than .05 and *** for p-values lower than .01.

Parameter	Idpt	Gauss	Gumbel	Frank
Pareto/NBD Component				
Transaction rate shape r	0.55^{***}	0.55^{***}	0.55^{***}	0.55^{***}
Transaction rate heterogeneity ρ	-2.36^{***}	-2.36^{***}	-2.36^{***}	-2.36^{***}
Dropout shape s	0.61^{***}	0.61^{***}	0.61^{***}	0.61^{***}
Dropout scale β	11.68^{***}	11.68^{***}	11.67^{***}	11.67^{***}
Gamma/Gamma Component				
Transaction value shape p	6.25^{***}	6.25^{***}	6.25^{***}	6.27^{***}
Heterogeneity shape q	3.74^{***}	3.74^{***}	3.76^{***}	3.76^{***}
Revenue rate heterogeneity κ	2.74^{***}	2.74^{***}	2.74^{***}	2.74^{***}
Association				
η_{inter}	0.00	0.41^{***}	-3.47^{**}	1.42^{***}

Table 2: Descriptive statistics of the securities transaction data.

	Mean	Std. Dev.	Minimum	Maximum
Number of repeated transactions	7.10	15.00	0.00	181.00
Recency *(in weeks)*	15.19	11.47	0.03	36.47
Average transaction value *(in €)*	2,677.08	2,073.93	26.27	1,8319.95
Age *(in years)*	50.92	15.83	18.00	80.00
Living area *(Dummy, City suburb = 1)*	0.22	0.42	0.00	1.00
Cohort *(in weeks)*	26.36	7.19	12.20	36.47
Primary bank *(Dummy)*	0.70	0.46	0.00	1.00

Table 3: Parameters estimates of the Gumbel model for the securities transaction data. Significance levels are indicated with * for p-values lower than .1, ** for p-values lower than .05 and *** for p-values lower than .01.

Pareto/NBD		**Gamma/Gamma**	
Transaction rate shape r	0.47^{***}	Transaction value shape p	6.69^{***}
Dropout shape s	0.20^{***}	Heterogeneity shape q	2.24^{***}
Dropout scale β	1.51^{***}		

Covariates	*Pareto/NBD*	*Gamma/Gamma*
Intercept	-0.24^{***}	223.74^{**}
Age	0.00	1.34^{*}
Living Area	-0.07^{**}	12.38^{***}
Cohort	-0.05^{***}	-0.51^{*}
Primary bank	0.25^{***}	42.07^{***}

Association	η_{inter}	η_{intra}
Intercept	-6.65^{***}	-3.48^{***}
Age	1.72^{***}	0.04^{***}
Living Area	0.43^{***}	-0.37^{***}
Cohort	-0.38^{***}	-0.02^{***}
Primary bank	-1.70^{***}	-0.64^{***}

Table 4: Number of customers incorrectly selected ("type II error") under the Gumbel and independent models for different threshold values, from the 99% to the 55% highest CLV percentile.

	Number of Incorrectly Selected Customers		
Percentile	Gumbel	Independent	Difference
99%	54	55	1
95%	156	163	7
90%	247	258	11
85%	278	297	19
80%	278	297	19
75%	301	320	19
70%	329	350	21
65%	363	387	24
60%	362	379	17
55%	341	362	21

Figure 1: Transaction values vs. interpurchase times of five imaginary customers. Customer averages are reported in bold.

Figure 2: Contours plots corresponding to the joint distribution of two standard-normal random variables with Spearman rank correlation equals to 50%, for (i) an independent copula (upper-left plot), (ii) a Gaussian copula (upper-right plot), (iii) a Gumbel copula (lower-left plot), and (iv) a Frank copula (lower-right plot).

Figure 3: Conditional histograms of the intra-customer Spearman's association conditioning on age, divided into four quartiles (left) and living area (right).

APPENDIX A: Copula Density Functions

Here, we list the probability density functions of the copula distributions used in this paper.

Gaussian copula A first family of copula that can be found in the literature is the Gaussian or normal copula, with density function

$$c_\theta(F,G) = \frac{1}{(1-\theta^2)^{1/2}} \exp\left(-\frac{1}{2}\psi'(R(\theta)^{-1} - I_2)\psi\right), \tag{13}$$

where $-1 < \theta < 1$, $\psi = (\Phi^{-1}(F), \Phi^{-1}(G))'$, Φ is the univariate standard normal distribution function, $R(\theta)$ is the correlation matrix

$$R(\theta) = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}, \tag{14}$$

and I_2 is the identity matrix of size 2. This copula permits both positive and negative association between the variables. Values of θ equal to -1, 0 and 1 correspond to the minimal value of negative association, independence, and the maximum of positive association. When combined with two normal marginal distributions, the joint distribution is bivariate normal. The Spearman correlation between random variables with a Gaussian copula distribution equals $\rho = \frac{6}{\pi}\arcsin(\theta/2)$.

Gumbel copula The density of the Gumbel (or logistic) copula is given by

$$c_\theta(F,G) = \frac{C(F,G)[\log(F)\log(G)]^{\theta-1}}{FG[(-\log(F))^\theta + (-\log(G))^\rho]^{2-1/\theta}} \left[\left((-\log(F))^\theta + (-\log(G))^\theta\right)^{1/\theta} + \theta - 1\right]. \tag{15}$$

where $1 \leq \theta < \infty$ is the assocation parameter. The copula distribution in (15) is

$$C_(F,G) = \exp\{-\left((-\log(F))^\theta + (-\log(G))^\theta\right)^{1/\theta}\}.$$

The limiting case $\theta = 1$ gives independence while for $\theta \to \infty$, one obtains a perfect dependency. The Gumbel copula is a special case of an Archimedean copula. There is no closed expression for the Spearman correlation as a function of θ.

Frank copula The density of the Frank Copula is given by

$$c_\theta(F,G) = \frac{\theta(1-e^{-\theta})e^{-\theta(F+G)}}{[(1-e^{-\theta}) - (1-e^{-\theta F})(1-e^{-\theta G})]^2}, \tag{16}$$

where $-\infty < \theta < \infty$. This copula permits both positive and negative association between the variables. Values of $-\infty$, 0 and ∞ correspond respectively to the smallest possible negative association, to the independent case, and to the largest possible positive association. The Spearman correlation between random variables with a Frank copula distribution is given by $\rho = 1 - \frac{12}{\theta}(D_2(-\theta) - D_1(-\theta))$, where D_k is the Debye function

$$D_k(\theta) = \frac{k}{\theta^k} \int_0^\theta \frac{t^\theta}{e^t - 1} dt.$$

APPENDIX B: The Copula-Extended CLV Model

A1: Interpurchase times $IPT_{i,j}$ are exponentially distributed with parameter λ_i.

A2: λ_i is gamma distributed with constant shape r and scale $\alpha_i = f(V_i)$, a function of the covariates V_i of customer i. We take $f(V_i) = \exp(-\rho' V_i)$

A3: Transaction values m_{ij} are gamma distributed with constant shape p, and scale parameter ν_i. The latter follow a gamma distribution with shape q and scale $\gamma_i = f(V_i)$. We take $f(V_i) = \exp(\kappa' V_i)$.

A4: The time to death of a customer is exponentially distributed with parameter μ_i.

A5: The parameter μ_i is gamma distributed with scale parameter β and shape parameter s, both constant over the population.

A6: The transaction rate λ_i and death rate μ_i are independent. The revenue rate ν_i and death rate μ_i are independent.

A7: The association between λ_i and ν_i is modeled by a copula c_{inter} with parameter $\theta_{inter,i} = f(\eta'_{inter} V_i)$.

A8: The association between m_{ij} and $IPT_{i,j}$ is modeled by a copula c_{intra} with parameter $\theta_{intra,i} = f(\eta'_{intra} V_i)$.

The (hyper)parameters of this model are $\boldsymbol{\theta} = (r, \rho, p, q, \kappa, \beta, s, \eta_{inter}, \eta_{intra})$

APPENDIX C: Prediction of CLV

Below, we outline how the CLV over the next H time units can be predicted from the estimated copula-extended model. We use the notations of Appendix B. For a given customer i with covariates V_i, we compute $\hat{\alpha}_i = \exp(-\hat{\rho}' V_i)$, $\hat{\gamma}_i = \exp(\hat{\kappa}' V_i)$, $\hat{\theta}_{inter,i} = f(\hat{V}_i' \eta_{inter})$, and $\hat{\theta}_{intra,i} = f(V_i' \hat{\eta}_{inter})$. Furthermore, we denote x_i the number of repeated transactions done by customer i, T_i the number of time units between his first transaction and the moment of prediction, and m_i the historical average of the transaction values. From these quantities, the probability p_i that the customer is still alive at the moment of prediction is computed using formulas (11)-(13) in Schmittlein et al. (1987). We use the expressions for the posterior distributions of the death, transaction and revenue rate derived in Schmittlein et al. (1987) and Fader et al. (2005b). A gamma distribution with shape parameter a and scale parameter b is denoted $gamma(a, b)$.

For every customer i, we generate $M = 3000$ values from the distribution of $CLV_{i,H}$, by simulating future transaction streams:

1. We draw a value from a uniform distribution on [0,1]. If this value is larger than p_i, we set $CLV^* = 0$ and consider the customer as "death." Otherwise, we continue to simulate the transaction process.

2. We draw a value (U_1^*, U_2^*) from the copula distribution with parameter $\hat{\theta}_{inter,i}$. Note that most software packages have build-in routines to do this for the copulas presented in Appendix A.

3. We compute λ^* as the inverse of the cdf of a $gamma(\hat{r} + x_i, \hat{\alpha}_i + T_i)$ distribution evaluated at U_1^*.

4. We compute ν^* as the inverse of the cdf of a $gamma(\hat{p}x_i + \hat{q}, \hat{\gamma}_i + m_i x_i)$ distribution evaluated at U_1^*.

5. We draw a value $\mu^* \sim gamma(\hat{s}, \hat{\beta} + T_i)$.

6. We draw the time of death τ^* from an exponential distribution with parameter μ^*.

7. Set $t^* = 0$ and $CLV^* = 0$. While $t^* \leq H$ and $t^* \leq \tau^*$

   (a) Draw a value (U_1^*, U_2^*) from the copula distribution with parameter $\hat{\theta}_{intra,i}$.

(b) Compute IPT^* as the inverse of an exponential cdf with parameter λ^* at U_1^*.

(c) Compute m^* as the inverse of a the cdf of a $gamma(\hat{p}, \nu^*)$ at U_2^*.

(d) Update $t^* \leftarrow t^* + IPT^*$.

(e) Update $CLV^* \leftarrow CLV^* + margin\,(m^*)\,(1+d)^{t^*}$.
Recall that d stands for the discount rate.

As such, we obtain M draws from the estimated distribution of $CLV_{i,H}$.