A Non-parametric Approach to Incorporating Incomplete Workouts Into Loss Given Default Estimates

Rapisarda, Grazia and Echeverry, David
Royal Bank of Scotland

16. November 2010

A Non-parametric Approach to Incorporating Incomplete Workouts Into Loss Given Default Estimates

Grazia Rapisarda and David Echeverry[‡]

November 16, 2010

Abstract

When estimating Loss Given Default (LGD) parameters using a workout approach, i.e. discounting cash flows over the workout period, the problem arises of how to take into account partial recoveries from incomplete workouts. The simplest approach would see LGD based on complete recovery profiles only. Whilst simple, this approach may lead to data selection bias, which may be at the basis of regulatory guidance requiring the assessment of the relevance of incomplete workouts to LGD estimation. Despite its importance, few academic contributions have covered this topic. We enhance this literature by developing a non-parametric estimator that - under certain distributional assumptions on the recovery profiles - aggregates complete and incomplete workout data to produce unbiased and more efficient estimates of mean LGD than those obtained from the estimator based on resolved cases only. Our estimator is appropriate in LGD estimation for wholesale portfolios, where the exposure-weighted LGD estimators available in the literature would not be applicable under Basel II regulatory guidance.

Keywords: Credit risk, bank loans, loss-given-default, LGD, incomplete observations, mortality curves.

JEL classification: C14, G32.

[*]grazia.rapisarda@rbs.com and david.echeverry@rbs.com

[†]The authors are Head of Modelling and Risk Manager at the Credit Portfolio Analytics Department of the Royal Bank of Scotland respectively. The points of view expressed in the document are those of the authors and do not represent those of the Royal Bank of Scotland or the Board of Directors. The authors are the only ones responsible for any error in the document.

Introduction

In its advanced internal rating based approach, the New Basel Accord (Basel II) requires credit institutions to calculate their capital requirements using their own estimates of Loss Given Default parameters (LGD). One of the most common approaches to estimating LGD is the workout approach. This is based on the estimation of individual default recovery rates obtained as the sum of discounted cash flows of recoveries over the length of the workout period. Information contained in individual recovery rates is then aggregated to produce LGD parameters that are assigned to individual facilities by means of parametric or non-parametric models based on facility and (or) borrower characteristics.

When using the workout approach, the problem arises of how to deal with partial recovery profiles of unresolved defaults. These observations may relate to either relatively recent defaults or older ones involved into particularly lengthy bankruptcy proceedings. The simplest approach would see these cases excluded from the estimation process altogether, with LGD based on complete recovery profiles only. Whilst simple, results based on this approach may be affected by data selection bias if unresolved cases contain information relevant to LGD which is not captured by the recovery profiles of resolved defaults. Even in cases where selection bias is not an issue, inclusion of partial workouts may still be relevant if they contribute to reduce the error around the estimates. This may be at the basis of some regulatory guidance (e.g. FSA guidance) requiring the inclusion of incomplete workouts in the LGD estimation process.

In this paper we present a non parametric estimator that aggregates complete and incomplete recovery profiles to produce unbiased estimates of LGD. Whilst building on the mortality rate approach to recovery rate estimation in the presence of incomplete workouts (Dermine and Neto (2005), Bastos (2009)), the method developed in this paper contributes to existing contributions in several respects as outlined below.

Our first contribution is a default-weighted reformulation of the exposure-weighted Kaplan-Meier estimator of ultimate recovery rates as presented in the literature. Under suitable discounting of future cash flows,[1] both formulations lead to the same measure of ultimate recovery rate when applied at individual loan level. Furthermore both formulations may lead to unbiased estimates; nevertheless the default-weighted one is viewed as more appropriate to ensure compliance with regulatory guidelines on LGD estimation for e.g. wholesale portfolios.

Incorporating incomplete workouts require a shift from "aggregate partial recoveries over time first and then across individuals" to "aggregate partial recoveries across individuals first and then over time", which is the principle of both methodologies. In the first case, ultimate recovery rates must be realisations

[1] For a suitable discounting in the multiplicative framework see Dermine and Neto (2005).

of the same random variable, whereas in the second recovery profiles need to be realisations of the same stochastic process. For instance, if the sample of unresolved cases has a large concentration of 'slow' workouts, as opposed to a larger concentration of 'quick' workouts in the sample of resolved cases, direct incorporation of unresolved cases into the estimates would bias the results. We draw attention to the consequences on the estimation output of not adequately capturing drivers of the difference in recovery dynamics when aggregating resolved and unresolved cases.

Finally we derive sufficient conditions on the distribution of recoveries over time for the estimator to lead to more precise LGD estimates than those based on resolved cases only. Changing the order of aggregation allows for incorporation of unresolved defaults but does not necessarily increase the precision of estimates. The efficiency gain with respect to the estimator based on resolved cases depends on the serial correlation of partial recoveries: in the presence of negative correlations (induced, for instance, by the fact that the ultimate LGD is not expected to be higher than 100%), the estimator could be less efficient. We determine a lower bound[2] on serial correlation that preserves the efficiency gain. The simplifying assumption of no serial correlation is widespread across the literature on net present value; we show that under such an assumption the proposed estimator is more efficient.

The paper is organized as follows. We review the existing literature and highlight this paper's contribution in section 1. In section 2 we introduce a statistical estimator of LGD based on both partial and complete recovery data and highlight its statistical properties. In section 3 we characterise the possible scenarios that arise when comparing the recovery dynamics of resolved and unresolved cases and identify the situations where aggregation of data across the two groups may improve LGD estimates. Section 4 concludes the paper by summarising the main results and suggesting possible areas of future research.

1 Literature review

Since the advent of the Basel Capital Accord, the modelling of recovery rates has received increasing interest by academics and practitioners. Various approaches have been developed, some leveraging on the option-valuation framework - already used to model probability of default parameters - others relying on historical data to derive estimates of LGD based on realised recovery rates. Whilst the majority of contributions within this latter strand rely on the exclusive use of resolved defaults data, a few have modelled LGD using both resolved and unresolved defaults (see Dermine and Neto (2005), Bastos (2009) and Moody's LossCalc (2006)). Our paper contributes to this literature by presenting a model for the estimation of LGD based on discounted recoveries of both resolved and

[2]Fuller and Kim (1980) derive the impact of serial correlation on the total variance of an estimator of discounted cash flows.

unresolved defaulted loans. Unresolved cases are incorporated by means of an estimator based on Kaplan and Meier (1958) and its study of mortality rates. It has been applied to estimate default rates (see Altman (2009) and Altman and Suggitt (2008)) but only recently to the estimation of recovery rates (see Dermine and Neto (2005), Bastos (2009)).

The Kaplan-Meier idea is as follows: marginal survival rates $m_{i,u}$ at a given time u over the workout period are obtained as cash flows during period u normalised by *discounted outstanding exposure* at the start of period u. Cumulative recovery rates at a given point in time t are then derived as the complement of the product of marginal recoveries, namely $M_{i,t} = 1 - \prod_{u=1}^{t}(1 - m_{i,u})$. Individual marginal recovery rates can then be aggregated regardless of whether they relate to resolved defaults or unresolved ones. Bastos (2009) and Dermine and Neto (2005) calculate average marginal recovery rates weighted by discounted loan outstanding at each point in time t across all defaults in the sample (resolved and unresolved) to come up with an estimate of the recovery rate.

Aggregation of the $m_{i,u}$ to produce an m_u can be achieved through default or exposure-weighting. While a default-weighted aggregation does not coincide with $\overline{RR}$, an exposure-weighted aggregation matches an exposure-weighted recovery rate. The second question is about discounting cash flows. Once the reference point moves from exposure at default to outstanding exposure, cash flows need to be discounted back to the start (or end) of the period and not to the time of default. Dermine and Neto (2005) provide an example under discrete discounting; for our purpose we consider a case of continuous discounting.

We develop a time-additive estimator of recovery rates. Defining the partial recovery rate at time u as the sum of cash flows over period u normalised by the *exposure at default*, the cumulative recovery rate is then the sum of the partial rates. At loan level this formulation coincides with the multiplicative version of the estimator. The proposed additive formulation yields an unbiased estimate, and is in fact a generalisation of the standard estimator which applies to resolved cases only. Furthermore we establish a condition on correlation of recoveries under which the estimator proposed is more efficient than the standard one.

The multiplicative version of Kaplan-Meier cannot be reconciled with the additive workout LGD estimator. Our first goal is to propose a non parametric default-weighted estimator of the recovery rate. The emphasis on default weighting is deemed more appropriate for use in LGD estimation for wholesale portfolios, where Basel II rules require benchmarking of LGDs to default-weighted averages of recovery rates. A model based solely on resolved cases is only preferable if the differences in dynamics do not imply differences in ultimate recoveries; controlling for different recovery dynamics is important to both the accuracy and precision of estimates. Our second goal is to provide a framework to assess

whether the inclusion of partial recovery data comes at the cost of biased or more volatile estimates.

2 An estimator of LGD based on both partial and complete recovery data

Incorporation of incomplete recovery data into LGD estimation implies introducing a new dimension into the analysis, namely recovery dynamics. Whilst by definition it is not possible to compare resolved and unresolved defaults on the basis of the ultimate recovery (and hence ultimate LGD), all default observations exhibit a partial recovery history. In what follows, an estimator is derived that will allow us to express recovery dynamics in terms of partial recovery at a given time after default. This will allow us to break down recovery histories into observation windows where they can be compared regardless of whether they relate to a complete or incomplete workout. Such is the idea behind the estimation of a mortality rate curve as outlined in Kaplan and Meier (1958).

For a given a facility i, we denote by $C_{i,t}$ the cash flow associated to it at time $t \in \{1, \ldots, T_i\}$ after default. T_i is the time up to which information on recovery cash flows[3] is available for facility i. Time is measured in discrete units which can be as granular as allowed by the data, and without loss of generality we will use annual intervals therein (what changes is what data you throw away for unresolved cases). Let R be the subset of the population where the full recovery history is observed. It follows that if $i \in R$ then T_i is the year where the last recovery cash flow is received (resolution year) and if $i \notin R$ then T_i coincides with the year preceding the year of truncation. We further denote by $O_{i,1}$ the amount of debt outstanding at default.[4]

For any loan i, let us denote with RR_i the ultimate recovery rate, i.e. the sum of discounted cash flows over the workout period

$$RR_i = \sum_{t=1}^{T_i} RR_{i,t}$$

where

$$RR_{i,t} = \frac{C_{i,t} e^{-rt}}{O_{i,1}}$$

denotes the (discounted) recovery at time t as a fraction of outstanding at default and r denotes the annual interest rate.[5]

[3] This may be a positive cash flow or confirmation that no cash flow has been received in the period.

[4] The amount outstanding at the start of period t $D_{i,t}$ is defined using discounted cash flows. An example is provided in Dermine and Neto (2005).

[5] Recent NPV literature has focused on the discounting of cash flows under serially correlated -stochastic- discount rates and uncorrelated -stochastic- cash flows. By assuming a constant interest rate we focus on the role of cash flows and their aggregation.

We denote the population of loans existing in period 1 by $P := P_1$. For a given $S \subseteq P$ we define $S_t = \{i \in S : \ t \leq T_i\}$ and $T_S = max\{T_i : \ i \in S\}$. For each $i \in R$ -where $R \subseteq P$ is the set of resolved loans- and $t \leq T_R$, RR_i is assumed to have mean ρ. In this case recovery information from resolved defaults is sufficient to produce an unbiased sample estimate of ρ based on the estimator

$$RR_S = \frac{\sum_{i \in S} RR_i}{|S|}$$

for every $S \subseteq R$. The standard estimator $\overline{RR} := RR_R$ is an unbiased estimator of ρ provided that $E[RR_i] = \rho$.[6] RR_i can only be observed on the set of resolved cases. By analogy with Kaplan and Meier (1958) we call it the reduced-sample estimate.

Remark 1. Let $\rho_t := E[RR_{i,t}]$ for each observation i. Then $\sum_{t=1}^{T_P} \rho_t = \rho$.

The observation is straightforward provided $T_P = T_R := T$ is the observation window length for all observations:

$$\sum_{t=1}^{T_P} \rho_t = \sum_{t=1}^{T_P} E[RR_{i,t}] = E\left[\sum_{t=1}^{T_P} RR_{i,t}\right] = E[RR_i] = \rho.$$

We aim to extend this estimator to allow inclusion of unresolved workouts. In order to do so, let us introduce some notation. Let $RR_{S,t} = \frac{\sum_{i \in S_t} RR_{i,t}}{|S_t|}$. We define the enlarged-sample estimate of ρ by

$$\widetilde{RR} := \sum_{t=1}^{T_S} RR_{S,t}.$$

Proposition 2. The enlarged-sample estimator is an extension of the reduced-sample one, that is $\widetilde{RR}_S = \overline{RR}_S$ for any $S \subseteq R$. In particular they coincide on the set R of resolved cases.

This result -proven in the appendix- relates the enlarged estimator to the reduced-sample one. Instead, the multiplicative Kaplan-Meier estimator coincides with an exposure-weighted average.

Proposition 3. For each $i \in P$ let $E[RR_{i,t}] = \rho_t$. The enlarged-sample estimator $\widetilde{RR}$ is then an unbiased estimator of ρ.

The proof is straightforward and can be found in the appendix. We have thus introduced the default-weighted statistic $\widetilde{RR}$ which under the given assumptions is an unbiased estimator of the recovery rate. We now provide a sufficient

[6] This implies that drivers of ultimate LGD on resolved cases (e.g. industry, region, loan type) are being controlled for.

condition for it to be more efficient than $\overline{RR}$, the average ultimate recovery rate on resolved cases. Recall that the efficiency of an estimator $\tilde{\theta}$ is defined by

$$e(\tilde{\theta}) = \frac{I_\theta^{-1}}{V_\theta[\tilde{\theta}]}$$

where $I_\theta^{-1} = E_\theta\left[\left(\frac{\partial \log f(x,\theta)}{\partial \theta}\right)^2\right]$ is the Fisher information amount, where $f(x,\theta)$ is the probability density of the random variable X.

It is plausible to assume that individual recovery profiles are independent, i.e. $Cov[RR_{i,s}, RR_{j,t}] = 0$ for each $1 \leq s \leq t \leq T$ if $i \neq j$. Serial correlation arises as past recoveries are likely to determine future ones. We assume a constant $Cov[RR_{i,s}, RR_{i,t}] = \sigma_{st}$ across individuals which yields the following:

Proposition 4. $e(\widetilde{RR}) \geq e(\overline{RR})$ if for each $t \in \{1, \ldots, T\}$ we have

$$\sum_{u=1}^{t}\{\sigma_{uu} + 2\sum_{s=u}^{T}\sigma_{us}\} \geq 0.$$

Notice that for $t = T$ the condition becomes the expression for $V[\widetilde{RR}]$ which is positive: adding a resolved observation to the sample always improves the efficiency of the estimator. Notice also that the condition is sufficient but not necessary: if for some t the inequality fails, the estimator might still be more efficient depending on the completeness of the observations added. The more data is provided per observation added, the less inequalities need to be met.

Different views may hold on what a suitable covariance structure would be. On one hand, strong partial recoveries are indicative of a successful workout with high recoveries; on the other hand, the exposure amount constitutes both an objective and a cap to the overall recovery, hence inducing a negative correlation among partial cash flows.

Different datasets might show different attributes. In the presence of serial correlation -a realistic assumption- proposition 4 provides a test on the covariance structure to determine which data increases the efficiency of the estimator. Current literature on net present value takes into account forms of dependence such as comonotonicity. Dhaene et al (2002 a, b) determine approximations for the distribution of the sum of comonotonic random variables when the distribution of the partial contributions is known but not the overall dependence structure.

3 Partial and complete recovery data: when is aggregation beneficial?

In this section we highlight the scenarios that arise when dealing with aggregation of partial and complete recovery data. This is based on the comparison between recovery dynamics of resolved and unresolved cases using the enlarged-sample estimator.

3.1 Preparing the data for estimation

We organise the observations into a $|P| \times T_P$ matrix with rows indicating the individual loan and columns indicating the number of years spent in recovery. For resolved cases, column entries coincide with actual (discounted) cash flows of recoveries at the time in which these have occurred, and a zero cash flow entry otherwise (including post-resolution years). One approach for unresolved cases is the following where no cash flow is observed a zero entry is input for years in recovery up to the last observation date (e.g. date at which the recovery profile was last updated) and a missing value for remaining (yet not observed) years in recovery.

Once recovery profiles are set, the enlarged-sample estimator $\widetilde{RR}$ can be applied to the data to obtain recovery rate profiles across groups of observations.

3.2 Analysing the recovery profiles

Figures 1 and 2 summarise possible dynamics of average partial recovery rates for resolved and unresolved defaults (i.e. $RR_{R,t}$ and $RR_{P\setminus R,t}$) where t is number of full years spent in recovery after default. We are interested in assessing the statistical significance[7] of the difference between these profiles and verify whether the conditions of proposition 3 for direct application of estimator $\widetilde{RR}$ are met.
Scenario 1 corresponds to the case of no statistical difference between average resolved and unresolved recovery profiles. This is the case described by proposition 3: direct aggregation of partial resolved and unresolved cases would be beneficial as evidence supports the assumption that recoveries of resolved and unresolved cases are random draws from the same distribution at each time t.

Scenario 2 illustrates the case where dynamics are significantly different for some time t and different hypotheses can be made with regard to the unobserved ultimate recovery rate of unresolved cases (once they resolve). These may end up resulting in an average recovery rate which is higher than that of resolved cases (case (a) in graph) or lower (case (c)) or about the same (case (b)). In all these cases, direct aggregation of partial and complete workouts by means of estimator $\widetilde{RR}$ would not be appropriate. Not only this would lead

[7] Statistical significance of the difference in means between $RR_{R,t}$ and $RR_{P\setminus R,t}$.

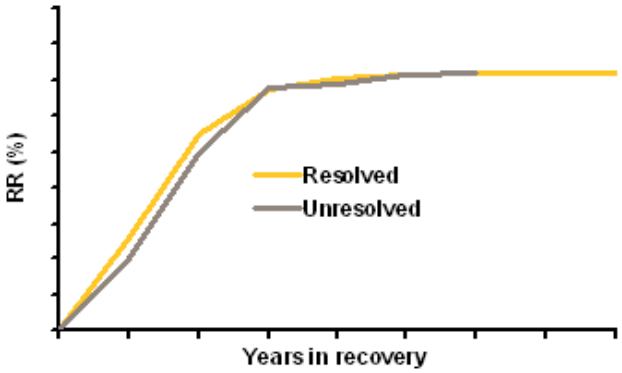

Figure 1: Scenario 1

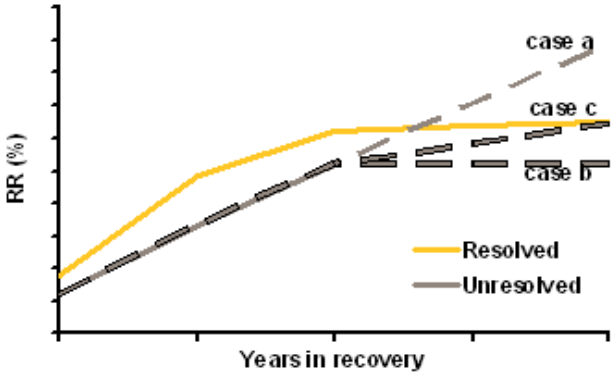

Figure 2: Scenario 2 cases a, b and c

to biased LGD estimates (upward or downward bias) but also to potentially unstable estimates as new recovery information is added to the data.[8]

In the following section we illustrate how to come up with unbiased and where possible more precise estimates of LGD by aggregating partial and complete recovery data. The approach to aggregation depends on the characteristics of the data and/or individual banks' preferences between a structural, parametric model versus a non parametric one. In what follows we will illustrate a non parametric approach to aggregating information to produce a non parametric model of LGD. For a comparison with alternative estimation approaches see section 1.

[8]Instability depends on the composition of recovery profiles of new unresolved cases added to the data.

3.3 Aggregating different recovery dynamics

Case (b) in figure 2 is particularly interesting: here recovery dynamics between resolved and unresolved cases differ but the enlarged-sample recovery rates are close. The factors driving the differences in partial rates may be different from those affecting ultimate rates, as differences in partial recovery rates may cancel out over time. In this case, LGD estimates based on resolved cases only would still be unbiased and justifiable on this basis. On the other hand, partial workouts information could be used to improve the precision of the estimator.

For simplicity, and without loss of generality, suppose there are only two different recovery dynamics. Let Y be a binary indicator variable available for all defaults that fully explains differences in partial recovery rates at a given time t. Let P^0 and P^1 be the set of defaults where $Y = 0$ and $Y = 1$ respectively. The factor Y explains the different partial recovery rates for both resolved and unresolved cases, i.e. $E[RR_{P^k,t}] = \rho_{k,t}$ for each $t \leq T_P$ and $k \in \{0, 1\}$. Suppose now that the two groups exhibit different recovery dynamics but the same ultimate recovery. This means for each t $E[RR_{i_k,t}] = \rho_{k,t}$ given $i_k \in P^k$ are such that $\sum_t \rho_{k,t} = \rho$ for each $k \in \{0, 1\}$. Then

$$\begin{aligned}
E[\widetilde{RR}] &= E\left[\sum_{t=1}^{T_P} RR_{P,t}\right] \\
&= \sum_{t=1}^{T_P} E[RR_{P,t}] \\
&= \sum_{t=1}^{T_P} \frac{\sum_{i\in P_t} E[RR_{i,t}]}{|P_t|} \\
&= \sum_{t=1}^{T_P} \frac{\sum_{i\in P_t^0} E[RR_{i,t}] + \sum_{i\in P_t^1} E[RR_{i,t}]}{|P_t|} \\
&= \sum_{t=1}^{T_P} \frac{|P_t^0|\rho_{0,t} + |P_t^1|\rho_{1,t}}{|P_t|}
\end{aligned}$$

which is different from ρ unless $\sum_{t=1}^{T_P} \rho_{0,t} = \sum_{t=1}^{T_P} \rho_{1,t} = \rho$ and $|P_t^0|$ remains constant over all t (which is in fact the case if $P = R$). However, in this case an unbiased estimator of ρ is given by

$$\alpha\widetilde{RR}_{P^0} + (1 - \alpha)\widetilde{RR}_{P^1}$$

for any $\alpha \in [0, 1]$. The estimator would still produce an unbiased estimate of the recovery rate.

Identification of the factors driving the differences between recovery dynamics may not be feasible, and even when possible the factors may not be easily measurable.

If no evidence to support conjecture (b) can be found, using information from resolved cases only would clearly lead to a biased LGD estimate. Whilst in scenario (a) the bias would be prudent and hence potentially acceptable from a regulatory point of view[9], scenario (c) would suggest that using resolved cases only would lead to non-conservative estimates of LGD and hence difficult to justify from a regulatory point of view. In both cases, using recovery information from both resolved and unresolved cases may lead to improved LGD estimates versus using resolved cases only.

Whether this result can be achieved depends on the extent to which the factors driving the differences in dynamics between the two groups can be identified with enough confidence and controlled for. Implementation of an LGD model based on this approach would thus require that the drivers be measurable on performing defaults as well.The estimation approach would in this case result in a more granular LGD model, with different LGD assigned to facilities/borrowers depending on the value of the variable Y.

4 Summary and conclusions

In this paper we have derived a statistical estimator that allows to incorporate incomplete workouts into LGD estimates. The estimator proposed here is more appropriate than exposure-weighted alternatives available in the literature for use in LGD estimation for wholesale portfolios, where the rules emphasise the need for benchmarking LGDs to default-weighted historical LGD rates. Both formulations arise from the Kaplan-Meier idea originally developed to estimate mortality rates in the presence of truncated samples. Under standard assumptions, the estimator leads to unbiased estimates of LGD. Efficient estimates are obtained under more stringent assumptions on the degree of correlation among cash flows recovered over the workout period.

'Resolvedness' is not an intrinsic feature of a given facility but a circumstantial one. For that reason, any empirical similarity or difference between resolved and unresolved cases must be decomposed into its drivers. If drivers can be controlled for, the enlarged sample estimator produces unbiased -and potentially more efficient estimates- of LGD parameters. Depending on whether such variables are also the drivers of ultimate LGD estimates will coincide with ultimate LGD estimates on resolved cases.

With regard to the efficiency property, model developers should consider that inclusion of unresolved cases may in fact increase rather than reduce the error around the estimates and as such may be considered not relevant to the estimation process.

[9] Observed cumulative recovery rates of unresolved cases actually cross resolved ones in the real data, implying estimates based on resolved cases only are conservative.

Further theoretical work would include further analysis of the asymptotic properties of the estimator with respect to the choice of the observation window (assumed to be annual in this paper). Also, in the multiplicative setting it is possible to derive a formula for the variance of the estimator, given that marginal recovery rates are independent. Further possible theoretical work could build on a comonotonic sum of partial recoveries to calculate the theoretical variance of the estimator proposed in this paper.

Appendix

Proof of proposition 2.

$$\begin{aligned}\widetilde{RR}_S &= \sum_{t=1}^{T_S} RR_{S,t} \\ &= \sum_{t=1}^{T_S} \frac{\sum_{i\in S_t} RR_{i,t}}{|S_t|}.\end{aligned}$$

For every $i \in S$, $i \in R$ and so if $t > T_i$ it follows $C_{i,t} = 0$. Thus $S_t = S$ for every $t \leq T_S$ and so

$$\begin{aligned}\widetilde{RR}_S &= \sum_{t=1}^{T_S} \frac{\sum_{i\in S} RR_{i,t}}{|S|} \\ &= \frac{\sum_{i\in S}\sum_{t=1}^{T_S} RR_{i,t}}{|S|} \\ &= \frac{\sum_{i\in S} RR_i}{|S|} \\ &= \overline{RR}_S.\end{aligned}$$

The multiplicative Kaplan-Meier estimator coincides with an exposure-weighted average under discounted outstanding amounts. For that we define the discounted outstanding $D_{i,t}$ defined recursively by $D_{i,1} = O_{i,1}$ and $D_{i,t+1} =$

$e^r D_{i,t} - C_{i,t}$

$$\begin{aligned}
M_S &= 1 - \prod_{t=1}^{T_S}(1 - m_{i,t}) \\
&= 1 - \prod_{t=1}^{T_S}\left(1 - \frac{\sum_{i\in S_t} C_{i,t}e^{-r}}{\sum_{i\in S} D_{i,t}}\right) \\
&= 1 - \prod_{t=1}^{T_S}\left(\frac{\sum_{i\in S_t} D_{i,t} - C_{i,t}e^{-r}}{\sum_{i\in S} D_{i,t}}\right) \\
&= 1 - \prod_{t=1}^{T_S}\left(\frac{\sum_{i\in S_t} D_{i,t+1}}{\sum_{i\in S} O_{i,t}}\right) \\
&= 1 - \frac{e^{-rT_S}\sum_{i\in S} D_{i,T_S+1}}{\sum_{i\in S} D_{i,1}} \text{ using the fact that } S_t \equiv S
\end{aligned}$$

Now $D_{i,t+1} = e^r D_{i,t} - C_{i,t}$ and so $D_{i,t+1} = e^{rt}D_1 - \sum_{u=1}^{t} C_u e^{r(t-u)}$, where $D_{i,1} = O_{i,1}$. Hence we conclude that

$$M_S = \frac{\sum_{i\in S}\sum_{t=1}^{T_S} C_{i,t}e^{-rt}}{\sum_{i\in S} O_{i,1}}$$

which is the exposure-weighted estimate of the recovery rate over the set $S \subseteq R$. □

Applying the same calculations to an average $\overline{m}_{i,t}$ instead of the exposure-weighted $m_{i,t}$ would lead neither to a default nor an exposure-weighted recovery rate.

Proof of proposition 3.

$$\begin{aligned} E[\widetilde{RR}] &= E\left[\sum_{t=1}^{T_P} RR_{P,t}\right] \\ &= \sum_{t=1}^{T_P} E\left[RR_{P,t}\right] \\ &= \sum_{t=1}^{T_P} \frac{\sum_{i\in P_t} E\left[RR_{i,t}\right]}{|P_t|} \\ &= \sum_{t=1}^{T_P} \frac{\sum_{i\in P_t} \rho_t}{|P_t|} \\ &= \sum_{t=1}^{T_P} \rho_t \frac{|P_t|}{|P_t|} \\ &= \rho \text{ by observation 1.} \end{aligned}$$

□

Proof of proposition 4. Since $E[\widetilde{RR}] = E[\overline{RR}]$, the Fisher information amount is the same for both estimators and so it is sufficient to prove that $V[\widetilde{RR}] \leq V[\overline{RR}]$. Under our assumptions on the covariance of partial recoveries, it can be proven that

$$Cov[RR_{P,s}, RR_{P,t}] = \frac{\sum_{i\in P_t} Cov[RR_{i,s}, RR_{i,t}]}{|P_s||P_t|}$$

which yields

$$V[\widetilde{RR}] = \sum_{t=1}^{T_P} \frac{\sigma_{tt} + 2\sum_{s>u} \sigma_{us}}{|P_t|}.$$

In order to determine the impact of an additional observation, let us call $x_t := |P_t|$. We then have

$$\begin{aligned} \frac{\partial V}{\partial x_t} &= \sum_{u=1}^{t} \frac{-1}{x_u^2}\left(\sigma_{uu} + 2\sum_{s=u}^{T_P} \sigma_{us}\right) \quad (1) \\ &= \frac{-1}{x_1^2}\left(\sigma_{uu} + 2\sum_{s=u}^{T_P} \sigma_{us}\right) \quad (2) \end{aligned}$$

For equation (1) notice that adding an observation at time t also adds an observation at times prior to t, i.e. $\partial x_s/\partial x_t = 1$ if $s < t$. Equation (2) follows from assuming without loss of generality that $x_u = x_1$ for $u \leq t$. From equation (2) it follows that $\partial V/\partial x_t \leq 0$ for every t if for all of them the inequality of proposition 4 holds. □

References

[1] Altman, Edward. *Measuring Corporate Bond Mortality and Performance.* The Journal of Finance, Vol. 44, No. 4. Sept. 1989.

[2] Altman, Edward. Default Recovery Rates and LGD in Credit Risk Modeling and Practise. 2009.

[3] Altman, Edward I., Resti, Andrea and Sironi, Andrea. *Default Recovery Rates: A Review of the Literature and Recent Empirical Evidence.* Journal of Finance Literature, 21-45. Winter 2006.

[4] Altman, Edward. and Suggitt, Heather. *Default Rates in the Syndicated Bank Loan Market: A Mortality Analysis.* Journal of Banking and Finance, Vol. 24, pp. 229-253. 2008

[5] Basel Committee on Banking Supervision, 2004. International Convergence on Capital Measurement and Capital Standards. Basel.

[6] Bastos, João. Forecasting bank loan loss-given-default. CEMAPRE. Technical University of Lisbon.

[7] Dermine, Jean and Neto de Carvalho, Cristina. *Bank Loan Losses-Given-Default: a Case Study.* Journal of Banking and Finance, Vol. 30, No. 4, pp. 1219-1243. Apr. 2006.

[8] Dhaene, Jan, Denuit, Michel, Goovaerts, Marc, Kaas, Rob and Vyncke, David. *The Concept of Comonotonicity in Actuarial Science and Finance: Applications.* Insurance: Mathematics and Economics, Vol. 31, No. 2, pp. 133-161. 2002

[9] Dhaene, Jan, Denuit, Michel, Goovaerts, Marc, Kaas, Rob and Vyncke, David. *The Concept of Comonotonicity in Actuarial Science and Finance: Theory.* Insurance: Mathematics and Economics, Vol. 31, No. 1, pp. 3-33. 2002

[10] Fuller, Russell J. and Kim, Sang-Hoon. *Inter-Temporal Correlation of Cash Flows and the Risk of Multi-Period Investment Projects.*The Journal of Financial and Quantitative Analysis, Vol. 15, No. 5, pp. 1149-1162. Dec. 1980.

[11] Kaplan, E. L. and Meier, Paul. *Nonparametric Estimation from Incomplete Observations.* Journal of the American Statistical Association, Vol. 53, No. 282, pp. 457-481. Jun. 1958.

[12] Moody's Investor Service, Special Comment. Measuring loss-given-default for structured finance securities: an update. Dec. 2006.

[13] Trück, Stefan, Harpaintner, Stefan and Rachev, Svetlozar. A Note on Forecasting Aggregate Recovery Rates with Macroeconomic Variables. Mar. 2005.