The impact of loan-to-value on the default rate of residential mortgage-backed securities
Luis Otero-González, Pablo Durán-Santomil, Rubén Lado-Sestayo and Milagros Vivel-Búa

Modeling joint defaults in correlation-sensitive instruments
Dariusz Gątarek and Juliusz Jabłecki

Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default
Han Sheng Sun and Zi Jin

Modeling corporate customers’ credit risk considering the ensemble approaches in multiclass classification: evidence from Iranian corporate credits
Parastoo Rafiee Vahid and Abbas Ahmadi
Risk.net in numbers

200+
New articles & technical papers (each month)

140,000
Users (each month)

370,000
Page views (each month)

58,000+
articles stretching back 20 years

6,500+
on Asset Management

6,900+
on Commodities

£21,000+
on Derivatives

19,400+
on Regulation

19,600+
on Risk Management

See what you’re missing
Visit the world’s leading source of exclusive in-depth news & analysis on risk management, derivatives and complex finance now.
The Journal of Credit Risk
EDITORIAL BOARD

Editors-in-Chief
ASHISH DEV Federal Reserve Board
MICHAEL GORDY Federal Reserve Board

Associate Editors
EDWARD ALTMAN NYU Stern School of Business
JENNIE BAI Georgetown University
J. L. BREEDEN Prescient Models LLC
JONATHAN CROOK The University of Edinburgh
DARRELL DUFFIE Stanford University
KAY GIESECKE Stanford University
GREG GUPTON Federal Reserve Board
JAY HUANG Penn State University
JOHN HULL University of Toronto
ROBERT JARROW Cornell University
AHMET KOCAGIL BlackRock
HOLGER KRAFT Goethe University
ANDRE LUCAS VU Amsterdam
DILIP MADAN University of Maryland
LORIANA PELIZZON Goethe University
DMITRY PUGACHEVSKY Quantifi
MICHAEL PYKHTIN Federal Reserve Board
DAN ROSEN S&P Capital IQ
PETER RITCHKEN Case Western Reserve University
JORGE R. SOBEHART Citi
STUART TURNBULL University of Houston
DONALD R. VAN DEVENTER Kamakura Corporation
FAN YU Claremont McKenna College
JING ZHANG Moody’s Analytics

SUBSCRIPTIONS

The Journal of Credit Risk (Print ISSN 1744-6619 | Online ISSN 1755-9723) is published quarterly by Incisive Risk Information Limited, Haymarket House, 28–29 Haymarket, London SW1Y 4RX, UK. Subscriptions are available on an annual basis, and the rates are set out in the table below.

<table>
<thead>
<tr>
<th></th>
<th>UK</th>
<th>Europe</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk.net Journals</td>
<td>£1945</td>
<td>€2795</td>
<td>$3095</td>
</tr>
<tr>
<td>Print</td>
<td>£735</td>
<td>€1035</td>
<td>$1215</td>
</tr>
<tr>
<td>Risk.net Premium</td>
<td>£2750</td>
<td>€3995</td>
<td>$4400</td>
</tr>
</tbody>
</table>

Academic discounts are available. Please enquire by using one of the contact methods below.

All prices include postage. All subscription orders, single/back issues orders, and changes of address should be sent to:

UK & Europe Office: Incisive Media (c/o CDS Global), Tower House, Sovereign Park, Market Harborough, Leicestershire, LE16 9EF, UK. Tel: 0870 787 6822 (UK), +44 (0)1858 438 421 (ROW); fax: +44 (0)1858 434958

US & Canada Office: Incisive Media, 55 Broad Street, 22nd Floor, New York, NY 10004, USA. Tel: +1 646 736 1888; fax: +1 646 390 6612

Asia & Pacific Office: Incisive Media, 20th Floor, Admiralty Centre, Tower 2, 18 Harcourt Road, Admiralty, Hong Kong. Tel: +852 3411 4888; fax: +852 3411 4811

Website: www.risk.net/journal  E-mail: incisivehv@subscription.co.uk

To subscribe to a Risk Journal visit Risk.net/subscribe or email corpsubs@risk.net
MANUSCRIPTS AND RESEARCH PAPERS SUBMITTED FOR CONSIDERATION MUST BE ORIGINAL WORK THAT IS NOT SIMULTANEOUSLY UNDER REVIEW FOR PUBLICATION IN ANOTHER JOURNAL OR OTHER PUBLICATION OUTLETS. ALL ARTICLES SUBMITTED FOR CONSIDERATION SHOULD FOLLOW STRICT ACADEMIC STANDARDS IN BOTH THEORETICAL CONTENT AND EMPIRICAL RESULTS. ARTICLES SHOULD BE OF INTEREST TO A BROAD AUDIENCE OF SOPHISTICATED PRACTITIONERS AND ACADEMICS.

SUBMITTED PAPERS SHOULD FOLLOW WEBSTER’S NEW COLLEGIATE DICTIONARY FOR SPELLING, AND THE CHICAGO MANUAL OF STYLE FOR PUNCTUATION AND OTHER POINTS OF STYLE, APART FROM A FEW MINOR EXCEPTIONS THAT CAN BE FOUND AT WWW.RISK.NET/JOURNAL. PAPERS SHOULD BE SUBMITTED ELECTRONICALLY VIA EMAIL TO: JOURNALS@INCISIVEMEDIA.COM. PLEASE CLEARLY INDICATE WHICH JOURNAL YOU ARE SUBMITTING TO.

YOU MUST SUBMIT TWO VERSIONS OF YOUR PAPER; A SINGLE LATEX VERSION AND A PDF FILE. LATEX FILES NEED TO HAVE AN EXPLICITLY CODED BIBLIOGRAPHY INCLUDED. ALL FILES MUST BE CLEARLY NAMED AND SAVED BY AUTHOR NAME AND DATE OF SUBMISSION. ALL FIGURES AND TABLES MUST BE INCLUDED IN THE MAIN PDF DOCUMENT AND ALSO SUBMITTED AS SEPARATE EDITABLE FILES AND BE CLEARLY NUMBERED.


THE NUMBER OF FIGURES AND TABLES INCLUDED IN A PAPER SHOULD BE KEPT TO A MINIMUM. FIGURES AND TABLES MUST BE INCLUDED IN THE MAIN PDF DOCUMENT AND ALSO SUBMITTED AS SEPARATE INDIVIDUAL EDITABLE FILES. FIGURES WILL APPEAR IN COLOR ONLINE, BUT WILL BE PRINTED IN BLACK AND WHITE. FOOTNOTES SHOULD BE USED SPARINGLY. IF FOOTNOTES ARE REQUIRED THEN THESE SHOULD BE INCLUDED AT THE END OF THE PAGE AND SHOULD BE NO MORE THAN TWO SENTENCES. APPENDIXES WILL BE PUBLISHED ONLINE AS SUPPLEMENTARY MATERIAL.

BEFORE SUBMITTING A PAPER, AUTHORS SHOULD CONSULT THE FULL AUTHOR GUIDELINES AT: HTTP://WWW.RISK.NET/STATIC/RISK-JOURNALS-SUBMISSION-GUIDELINES

QUERYS MAY ALSO BE SENT TO:
THE JOURNAL OF CREDIT RISK, INCISIVE MEDIA,
HAYMARKET HOUSE, 28–29 HAYMARKET, LONDON SW1Y 4RX, UK
TEL: +44 (0)20 7004 7531; FAX: +44 (0)20 7484 9758
E-MAIL: JOURNALS@INCISIVEMEDIA.COM

TO SUBSCRIBE TO A RISK JOURNAL VISIT RISK.NET/SUBSCRIBE OR EMAIL CORPSUBS@RISK.NET
The Journal of Credit Risk

With the rewriting of the Basel accords in international banking and their ensuing application, interest in credit risk has never been greater. The Journal of Credit Risk is at the forefront in tackling the many issues and challenges posed by the recent financial crisis, focusing on the measurement and management of credit risk, the valuation and hedging of credit products, and the promotion of greater understanding in the area of credit risk theory and practice.

The Journal of Credit Risk considers submissions in the form of research papers and technical reports on, but not limited to, the following topics.

- Modeling and management of portfolio credit risk.
- Recent advances in parameterizing credit risk models: default probability estimation, copulas and credit risk correlation, recoveries and loss given default, collateral valuation, loss distributions and extreme events.
- The pricing and hedging of credit derivatives.
- Structured credit products and securitizations, eg, collateralized debt obligations, synthetic securitizations, credit baskets, etc.
- Measuring, managing and hedging counterparty credit risk.
- Credit risk transfer techniques.
- Liquidity risk and extreme credit events.
- Regulatory issues, such as Basel II, internal ratings systems, credit-scoring techniques and credit risk capital adequacy.

To subscribe to a Risk Journal visit Risk.net/subscribe or email corpsubs@risk.net
CONTENTS

TECHNICAL REPORT

The impact of loan-to-value on the default rate of residential mortgage-backed securities 1
Luis Otero-González, Pablo Durán-Santomil, Rubén Lado-Sestayo and Milagros Vivel-Búa

RESEARCH PAPERS

Modeling joint defaults in correlation-sensitive instruments 15
Dariusz Gatarek and Juliusz Jablecki

Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default 43
Han Sheng Sun and Zi Jin

Modeling corporate customers’ credit risk considering the ensemble approaches in multiclass classification: evidence from Iranian corporate credits 71
Parastoo Rafiee Vahid and Abbas Ahmadi

Editors-in-Chief: Ashish Dev, Michael Gordy
Publisher: Nick Carver
Journals Manager: Dawn Hunter
Editorial Assistant: Carolyn Moclair

Subscription Sales Manager: Aaraa Javed
Global Key Account Sales Director: Michelle Godwin
Composition and copyediting: T&T Productions Ltd
Printed in UK by Printondemand-Worldwide

©Copyright Incisive Risk Information (IP) Limited, 2016. All rights reserved. No parts of this publication may be reproduced, stored in or introduced into any retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the copyright owners.
Technical Report

The impact of loan-to-value on the default rate of residential mortgage-backed securities

Luis Otero-González, Pablo Durán-Santomil, Rubén Lado-Sestayo and Milagros Vivel-Búa

Facultad de Economicas e Empresariais, Universidad de Santiago de Compostela, Avenida do Burgo, s/n. Campus Norte, 15782 Santiago de Compostela, Spain;
emails: luis.otero@usc.es; pablo.duran@usc.es; ruben.lado@usc.es; mila.vivel@usc.es

(Received February 10, 2015; revised February 18, 2016; accepted February 22, 2016)

ABSTRACT

This paper analyzes the validity of using the loan-to-value (LTV) ratio to explain the behavior of mortgage borrowers at an empirical level. To perform this analysis we use data for mortgage loan portfolios securitized in Spain during the period 2005–8. In the regression models developed, we find that higher initial LTV ratios are associated with greater default risk. The relation between the probability of default and LTV seems to be nonlinear, and a sharp increase is seen for values greater than 80%. Our findings confirm the adequacy of the new Basel III proposal that sets nonlinear capital requirement levels for banks holding residential mortgage loans at different LTV ratios. However, the significance shown in the regression models estimated with the “seasoning” variable could be considered in order to improve the models used to measure capital requirements.

Keywords: securitization; subprime; loan-to-value; default; conditional probability.
1 INTRODUCTION

Predicting the default rate is one of the main elements of the credit risk management process. In the particular case of mortgages, the probability of default is related to the loan-to-value (LTV), among other factors. The relationship between the two variables could be explained by the equity theory of default (Deng et al 2000), where borrowers base their default decisions on a rational comparison between returns and costs related to continuing to make mortgage payments. Under this theory, the borrowers have a put option with a strike equal to the value of the loan, and they will exercise it when the asset value drops below the value of the loan. The relation between the probability of default and LTV seems to be nonlinear and there is a sharp increase for an LTV greater than 80% (Campbell and Cocco 2015).

These aspects were still under discussion at the time of writing, during the implementation of the Basel III capital standards, with a proposal to set nonlinear capital requirement levels for banks holding residential mortgage loans at different LTV ratios. Regulatory capital requirements are designed to guarantee that banks have a sufficient quantity of capital to absorb losses during financial crises. In the rollout of Basel II, capital requirements for mortgage exposures were harmonized across the European Union at a 35% risk weight (RW). However, under Directive 2006/48/EC, competent authorities may tighten the capital requirements of these mortgage exposures by setting a higher RW (European Parliament 2006, Annex VI, Paragraph 9.1). Some European countries use LTV as a means of establishing RW, so mortgages above a certain LTV ratio must have an additional capital charge. The Basel Committee on Banking Supervision (BCBS) is aware that a 35% RW approach lacks risk sensitivity, so in recent a consultative document (Basel Committee on Banking Supervision 2015) it proposed to introduce a table of RWs ranging from 25% to 100%, based on the LTV ratio, because it believes that the LTV ratio is the most appropriate risk driver.

In this regard, it seems important to analyze the performance of the LTV approach empirically. To perform this analysis we use the data for mortgage loan portfolios securitized in Spain between 2004 and 2008. The Spanish securitization market was one of the most important in Europe during this period. According to data published by the European Securitisation Forum,¹ the volume of Spanish residential mortgage-backed securities (RMBSs) issues accounted for 14.88% of the total volume of European RMBSs issues in 2006, 18.49% in 2007, 10.43% in 2008 and 14.48% in 2009. Taking into account the outstanding balance of mortgage-backed securities at the end of the third quarter of 2009, Spain was in third place with a total of €167.1 billion (14.48% of the total), behind the United Kingdom (with €458 billion and 39.67%.

of the total) and Holland (with €202.4 billion and 17.53% of the total). Moreover, during that period, based on data provided by the INE (Spanish Statistical Office), prices at the end of 2010 accounted for a 13% drop. This is an ideal scenario setup for the analysis of equity default theory. Moreover, in Spain most of the mortgage loans do not limit the liability of the debtor to the value of their home. In case of default, the mortgage is foreclosed, and if the proceeds are not equivalent to the value of the debt, the lender may proceed against the remaining assets of the debtor. In the Spanish case this implies that the option is on all the debtor’s assets, not just on the value of the home, which could affect the performance of the equity default theory.

The aim of this paper is to provide empirical evidence on default behavior for the Spanish RMBS market. This work contributes to the existing literature by presenting new evidence of the relation between LTV and RMBS default rate. Note that, despite the importance of the issue, as far as we are aware there are few empirical studies in the literature and most of them do not refer to this particular subject. Our results will be useful for regulators, as they provide further insight into the amount of capital that banks should hold against residential mortgage portfolios with different LTV profiles. We also believe the analysis can be used by the research and investor communities to assess the risk of default caused by negative equity.

The paper is structured as follows. Section 2 reviews prior research. Section 3 details the empirical analysis. Finally, Section 4 presents the conclusions of the study.

2 LITERATURE REVIEW

In many studies on mortgage default risk, the LTV ratio is highlighted as the most important variable in determining the likelihood of mortgage loan default (Otero et al 2015). Under the equity theory of default, borrowers base their default decisions on a rational comparison of returns and costs related with ongoing mortgage payments. From the perspective of option pricing (Archer et al 2002), the mortgage borrower is considered to have a property sale option equivalent to the value of the mortgage principal. Therefore, the greater the LTV, the higher the intrinsic value of the option, the more likely it is to be exercised and the greater the incentive to incur default. Under this theory, the LTV ratio is considered to be the most important factor in default decisions (Wong et al 2004). As shown by Wong et al, the default probability is a nonlinear function of the independent variables, where LTV is included. According to the model proposed by Deng et al (2000), the option to default is exercised when the value of the property drops below the strike (value of the loan) by some specific amount. The default is not automatic when the equity value becomes negative because borrowers prefer to wait until default is irreversible (Foster and Van Order 1985). In

2 See http://www.ine.es.
this regard, Kau et al (1994) show that it may be optimal to wait to default until the house price is as much as 15% below the mortgage value. In addition, Deng and Quigley (2012) point out that higher initial LTV ratios are associated with greater default risk, since borrowers who have less wealth available for a down payment are more likely to be constrained. This relationship has been verified empirically by Vandell (1978), Campbell and Dietrich (1983), Schwartz and Torous (2003) and Mayer et al (2009). The model proposed by Campbell and Cocco (2015) also establishes that a small down payment reduces the incentive of borrowers to continue meeting their payments. The relation obtained between the probability of default and LTV in the model seems to be nonlinear, and the probability increases excessively for values greater than 90%. Although this model is characteristic of the US market, where in many states the mortgage is backed purely by the value of the collateral as there is no universal recourse to the borrower’s other assets, in the Spanish market there is also a direct relationship between the likelihood of default and the LTV. This is because the value of the property has a direct effect on the borrower’s willingness to pay off the mortgage, which drops as the value of the property decreases.

Despite the importance of LTV in default decisions, another important factor is related to the economic capacity of the borrower to pay the mortgage. In addition, there are other economic and social factors for avoiding default, such as: switching the costs of a new home; the legal environment; the mobility of those not considered in the call option view. The ability-to-pay hypothesis of default (the cashflow approach) states that mortgagors will always try to pay off the loan if their income flows are sufficient to meet the periodic payment without undue financial burden. Deng et al (2000), Fisher (2005) and Foote et al (2008) support the ability-to-pay theory. Finally, as Campbell and Cocco (2015) state, “regulators should think about combinations of LTV and LTI [loan-to-income], and should not try to control these parameters in isolation”.

### 3 EMPIRICAL ANALYSIS

We analyzed empirically the impact of LTV on the probability of default of the Spanish market RMBSs. To this end, we created a database of the characteristic elements of each of the securitization issues in Spain from 2005 to 2008. With this analysis, comprised of 138 outstanding issues, we considered 97.9% of the mortgage-backed securities market. The information sources consulted were the prospectuses of these issues and presale reports published by the respective rating agencies. We also used Bloomberg to obtain a series of related data on the evolution of the transaction, such

---

3 According to the volume of outstanding securitizations in Spain published by the European Securitisation Forum.
The impact of loan-to-value on the default rate of residential mortgage-backed securities

FIGURE 1 Loan-to-value and default rate for the overall sample.

![Graph showing the relationship between loan-to-value and default rate.]

Source: Bloomberg and compilation by the author. The initial weighted average LTV of each securitization is plotted on the horizontal axis, and the final default rate calculated at the end of 2010 is shown on the vertical axis. The higher the initial weighted LTV, the greater the default rate.

As date of issue or level of default of the loan portfolio, as well as another series of variables. Based on this information, we analyze the importance of the equity theory of default in explaining the default rate in the Spanish RMBS market. We also believe that it is necessary to include information on the average scoring of the credits included in the portfolio (although this data is not available from our database) in order to test the ability-to-pay theory, and therefore the effect of income and the interaction between both variables.

3.1 Dependent variable

As the dependent variable we use the default rate of each securitization at the end of 2010, provided by Bloomberg. As can be seen in Figures 1 and 2, there is a relation between the rate of default and the LTV calculated at the origination date of each issue. Furthermore, this relation seems to be nonlinear and, with some exceptions, there is a sharp increase for levels of between 80% and 100%.

This relation holds even if we differentiate by year. The dispersion around a certain level of LTV is explained by other characteristics related to the debtors’ ability and inherent willingness to pay. Obviously, the ability of banks to select better customers could be reflected in the level of default given a certain LTV. Figures 1 and 2 also show that there is lot of dispersion, especially around WALTV = 0.7. This is possibly...
due to other characteristics that affect the probability of default (size, volume, credit score, etc), many of which will be controlled for later.

### 3.2 Hypothesis and variables

The explanatory variable is the weighted average LTV of each securitization at the moment of issue, defined as the initial weighted average LTV (WALTV), and we establish the following hypotheses for the relation between LTV and the probability of default based on the previous theoretical analysis.

**Hypothesis 1:** a higher LTV is considered to have a positive effect on the default rate of mortgage-backed securities.

**Hypothesis 2:** the relation between LTV and default rate is nonlinear.

In addition, we have included some control variables that could affect the default rate. In this sense, we include the age in months (or months-on-books (MOB)) of the mortgage loans included in the securitization transactions at issue (WASeasoning). Seasoned loans are considered less likely to default because the borrowers have demonstrated their ability to pay (DBRS 2011). Glennon and Nigro (2011) showed that
TABLE 1  Variables and hypotheses considered in the study.

<table>
<thead>
<tr>
<th>Determinants</th>
<th>Variable</th>
<th>Prediction</th>
<th>Definition</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>LTV</td>
<td>Weighted average LTV (WALTVD)</td>
<td>+</td>
<td>Initial weighted average LTV</td>
<td>The higher the initial weighted LTV, the greater the credit risk</td>
</tr>
<tr>
<td>Seasoning</td>
<td>Average age (WASeasoning)</td>
<td>−</td>
<td>Average age of loans included in the portfolio (in months)</td>
<td>The greater the age of the portfolio, the lower the likelihood of default, and therefore the lower the likelihood of being classified in the subprime group</td>
</tr>
<tr>
<td>Diversification</td>
<td>Initial securitized volume (Volinitial)</td>
<td>−</td>
<td>Total securitized volume</td>
<td>The larger the securitized volume, the greater the diversification of the loan portfolio, and therefore the lower the risk</td>
</tr>
<tr>
<td>Loan size</td>
<td>Average loan value (Loanaverage)</td>
<td>+</td>
<td>Average amount of securitized loan portfolio</td>
<td>The larger the amount, the higher the default rate</td>
</tr>
</tbody>
</table>

seasoning is a key factor in making accurate forecasts of losses in mortgage portfolios. There is also evidence of a negative impact on performance in portfolios with a high credit concentration (Rossi et al 2009). Thus, diversification seems to be an important determinant of default rate. In this sense, we include the initial securitized amount of each securitization (Volinitial) as a proxy for the diversification effect associated with the size of the securitization. In addition, larger loan mortgages, defined as the average amount of securitized loan portfolio (Loanaverage), have higher default rates because expensive properties are more illiquid and take longer to sell during a market downturn (Qi and Yang 2009).
TABLE 2  Descriptive statistics of continuous independent variables.

<table>
<thead>
<tr>
<th></th>
<th>Observations</th>
<th>Mean</th>
<th>Standard deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>WALTV</td>
<td>138</td>
<td>0.6997</td>
<td>0.1135</td>
<td>0.4616</td>
<td>0.9533</td>
</tr>
<tr>
<td>WASeasoning</td>
<td>138</td>
<td>22.31</td>
<td>8.65</td>
<td>7.8</td>
<td>57</td>
</tr>
<tr>
<td>Loanaverage</td>
<td>137</td>
<td>133.628</td>
<td>37 426.01</td>
<td>22 177</td>
<td>222 576</td>
</tr>
<tr>
<td>Loaninitialvol</td>
<td>138</td>
<td>20.73</td>
<td>0.79</td>
<td>18.24</td>
<td>22.86</td>
</tr>
</tbody>
</table>

This table presents summary statistics for the variables used in the analysis. The sample includes 138 outstanding issues, representing 97.9% of the outstanding balance of mortgage-backed securities at the end of 2009. The data sources consulted were the prospectuses of these issues and presale reports published by the respective rating agencies. We also used Bloomberg to obtain a series of related data on the evolution of the transaction, such as date of issue or level of default of the loan portfolio, as well as another series of variables. “WASeasoning” is the average age of the loans. “Loanaverage” is the average size of the loans included in the portfolio. “Loaninitialvol” is the log of the value of the securitization.

TABLE 3  Description of the securitization sample.

<table>
<thead>
<tr>
<th>Year of issue</th>
<th>Initial securitized volume (€m)</th>
<th>Average LTV (%)</th>
<th>Average seasoning (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>29 164</td>
<td>66.33</td>
<td>20.16</td>
</tr>
<tr>
<td>2006</td>
<td>38 929</td>
<td>69.00</td>
<td>20.59</td>
</tr>
<tr>
<td>2007</td>
<td>62 055</td>
<td>73.64</td>
<td>21.44</td>
</tr>
<tr>
<td>2008</td>
<td>62 593</td>
<td>70.41</td>
<td>25.16</td>
</tr>
</tbody>
</table>

3.3 Descriptive analysis

Table 2 shows the main descriptive statistics of the variables used in the empirical multivariable analysis. Our sample is composed of 138 issues by the Spanish RMBS market from 2005 to 2008, representing almost the entire volume issued during the period under review. Thus, the securitization transactions in the sample have an average loan of €133 628 and the weighted average age is 22.31 months. Finally, the WALTV of the securitization transactions is 69.97%; 23.05% of the loans have LTV levels over 80%; and 0.54% of the loans have LTV levels over 100% on average.

Table 3 shows a more detailed description of the total sample regarding initial LTV levels and seasoning according to the year of the transaction origination.

Table 4 shows the correlation matrix of the independent variables used in the empirical analysis. As can be seen, there are significant correlations between the expected sign and the dependent variable for a large number of independent variables. The highest correlations are between WALTV and default rate (with a value of 0.40) and between WALTV and the average loan amount (with a value of 0.54), both with a positive sign.
### 3.4 Explanatory models of default rate

Our goal was to create a model to explain the empirical relation between WALT V and the rate of default. In addition, we check the importance of LTV in predicting subprime securitizations. We create a multiple regression model in which the dependent variable is determined by the default rate logarithm, while the main explanatory variable is the loan-to-value:

$$\log(PD)_i = c_i + \beta_{1i} \text{WALT V} + \sum_{j=1}^{k} \beta_{j+1,i} C_j + \varepsilon_i,$$

where $\log(PD)_i$ is the natural logarithm of default rate estimated for each regression $i$ (years 2006, 2007, 2008 and period 2005–8), $c_i$ is the constant term for each regression, WALT V is the explanatory variable, $C_j$ are the control variables (Loaninitialvol, WASeasoning and Loanaverage), $\beta$ is a matrix of parameters to be estimated (four parameters for each regression) and $\varepsilon_i$ is the error term of each regression.

The estimated models have a good fit to the data (Table 5), with an average $R^2$ around 0.45. In addition, the most important explanatory factor is the WALT V, and the relation is nonlinear; this supports both hypotheses tested in our study. In any case, we consider that the incorporation of the LTI or the borrower scoring would considerably increase the explanatory power of the model. Moreover, the $F$-statistic is very significant, indicating the validity of the proposed models. The validity of the model presented is also based on the absence of heteroscedasticity problems, contrasted by the White test, which is highly significant in all cases. Finally, the value of VIF shows no problems of multicollinearity between the variables analyzed. The fact that the models estimated for each year and for the entire period exceed the test discussed above justifies their validity for checking the hypotheses.

Table 6 shows the $R$-squared value of the estimated models that exclusively consider LTV as the explanatory variable and subsequently incorporate the effect of the “seasoning” variable. As we can see, the first models show the LTV is a very important

---

**Table 4** Table of correlations.

<table>
<thead>
<tr>
<th></th>
<th>Default rate</th>
<th>WALT V</th>
<th>WASeasoning</th>
<th>Loanaverage</th>
<th>Loaninitialvol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Default rate</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WALT V</td>
<td>0.4012</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WASeasoning</td>
<td>-0.4285</td>
<td>-0.1645</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Loanaverage</td>
<td>0.3541</td>
<td>0.5474</td>
<td>-0.4385</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Loaninitialvol</td>
<td>0.0328</td>
<td>-0.1022</td>
<td>-0.1469</td>
<td>0.0296</td>
<td>1</td>
</tr>
</tbody>
</table>

Table shows Pearson’s correlation coefficients for variables.
TABLE 5  Multiple regression estimates of Spanish RMBS default.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WALTIV</td>
<td>5.617**</td>
<td>4.460**</td>
<td>1.528*</td>
<td>4.139**</td>
</tr>
<tr>
<td>Loaninitialvol</td>
<td>−0.174</td>
<td>0.089</td>
<td>−0.223</td>
<td>−0.025</td>
</tr>
<tr>
<td>WASSeasoning</td>
<td>−0.081**</td>
<td>−0.063**</td>
<td>−0.075**</td>
<td>−0.006**</td>
</tr>
<tr>
<td>Loanaverage</td>
<td>−0.884</td>
<td>−0.782</td>
<td>0.051</td>
<td>−0.211</td>
</tr>
<tr>
<td>C</td>
<td>8.036</td>
<td>20.323</td>
<td>0.567</td>
<td>−2.473</td>
</tr>
<tr>
<td>Observations</td>
<td>32</td>
<td>42</td>
<td>42</td>
<td>137</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.475</td>
<td>0.4344</td>
<td>0.563</td>
<td>0.401</td>
</tr>
<tr>
<td>$F$-statistic</td>
<td>7.600**</td>
<td>8.940**</td>
<td>13.110**</td>
<td>24.460**</td>
</tr>
<tr>
<td>White test</td>
<td>10.250**</td>
<td>14.870**</td>
<td>10.960**</td>
<td>22.150**</td>
</tr>
<tr>
<td>VIF</td>
<td>1.71</td>
<td>1.77</td>
<td>1.50</td>
<td>1.36</td>
</tr>
</tbody>
</table>

This table includes the regression that estimates the relation between the default rate logarithm and the independent variables mentioned. * and ** denote statistical significance at the 5% and 1% confidence levels, respectively. For each model in turn, we checked the lack of multicollinearity between the variables by calculating the variance increase factor (VIF): to ensure that the models have no multicollinearity problems, they should have a VIF below 10. We also include White test data for heteroscedasticity.

TABLE 6  Level of fit measured by $R$-squared.

<table>
<thead>
<tr>
<th></th>
<th>LTV model</th>
<th>LTV and seasoning models</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>0.2013</td>
<td>0.4485</td>
</tr>
<tr>
<td>2007</td>
<td>0.3039</td>
<td>0.3995</td>
</tr>
<tr>
<td>2008</td>
<td>0.1949</td>
<td>0.5319</td>
</tr>
<tr>
<td>(2005–2008)</td>
<td>0.2104</td>
<td>0.3923</td>
</tr>
</tbody>
</table>

determinant of default risk. However, the explanatory power greatly improves when seasoning is incorporated. This means that regulation of capital requirements based on the LTV must be complemented by other factors, such as seasoning or LTI.

Regarding the results obtained in our regression models, we first highlight that the WALTIV ratio was significant, with a high level of confidence in most models, and it has a positive sign, producing the expected effect that a higher LTV ratio would increase the default probability of Spanish mortgagors (see Figures 3 and 4). These results support those of Schwartz and Torous (2003), Deng and Quigley (2012) and Campbell and Cocco (2015) and are consistent with the equity value theory; they show the sensitivity of Spanish debtors’ housing values relative to their debt. Undoubtedly, falling housing prices in Spain affected the default rates recorded in 2010; these have since soared to unprecedented levels. The projection of the default rate from the regression models, using only the LTV as an explanatory variable, shows how the
relationship is clearly nonlinear, and the default rate soars from an LTV of around 80%. The results are confirmed for the models fitted for each year and for the global model, although for 2006 and 2007 we can observe a higher gradient, associated with further relaxation of standards when granting loans. Our findings are also consistent with those made by Campbell and Cocco (2015) and in line with the BCBS’s proposal explained in Section 1.

The estimations of default rates adding the seasoning shows a moderating effect that is greater with higher levels of seasoning. This result supports the need to combine
LTV with other variables, in particular seasoning, LTI or credit scoring. Although LTI and credit scoring were unavailable in our database, Deng et al (2000) and Foote et al (2008) support the ability to pay as an important determinant of default probability.

4 CONCLUSIONS

This paper analyzed the validity of the loan-to-value ratio to explain the behavior of mortgage borrowers in Spain and to establish capital requirements. Using the regression models developed, we showed that a higher LTV ratio increases the default probability of Spanish mortgagors. The results are confirmed for the models fitted for each year and for the global model, although for the 2006 and 2007 observed a higher gradient, associated with a greater relaxation of standards when granting loans. Our results support those of Schwartz and Torous (2003), Deng and Quigley (2012) and Campbell and Cocco (2015), who found that higher initial LTV ratios are associated with greater default risk. The projection of the default rate from the regression models showed how the relationship is clearly nonlinear and the default rate soars from an LTV around 80%. Our findings are also consistent with those made by Campbell and Cocco (2015) and confirm the adequacy of the reform proposed in Basel III establishing nonlinear capital requirements that depend on the LTV of the portfolio. However, the significance of other variables, eg, seasoning or LTI, could be considered in order to refine the models used to estimate the probability of default and to determine capital requirements. In addition, our findings are consistent with the equity default theory and show the sensitivity of Spanish debtors’ housing values to debt. This result confirms the validity of default equity theory for countries with full recourse mortgage loans. In our view, the reason there are no major differences is because debtors with high LTVs are the most vulnerable and have little to lose in the case of default.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES


Get the information you need straight to your inbox

CHOOSE YOUR PREFERENCES

- Asset Management
- Commodities
- Derivatives
- Regulation
- Risk Management

We also offer daily, weekly and live news updates

Visit Risk.net/alerts now and select your preferences

To subscribe to a Risk Journal visit Risk.net/subscribe or email corpsubs@risk.net
Research Paper

Modeling joint defaults in correlation-sensitive instruments

Dariusz Gałtarek and Juliusz Jabłecki

1Polish Academy of Sciences, Al. 3 Maja 2, m. 164, 00-391 Warsaw, Poland; email: dgaltarek@yahoo.co.uk
2Faculty of Economic Sciences, Warsaw University, Długa 44/50, 00-241 Warsaw, Poland
3Economic Institute, National Bank of Poland, Świętokrzyska 11/21, 00-919 Warsaw, Poland; email: juliusz.jablecki@nbp.pl

(Received April 13, 2015; revised February 3, 2016; accepted February 29, 2016)

ABSTRACT

This paper presents a simple model for joint defaults and shows how it can be applied to pricing and risk-managing instruments that are sensitive to credit correlation, from simple repos to collateralized debt obligations. The model relies on a conservative and intuitive representation of a systematic factor as a chain of dependencies running through the whole economy. This allows capturing the concentration of defaults in time and endogenously produces dynamics of default correlation as the model output rather than its input.

Keywords: default correlation; Marshall–Olkin; European Financial Stability Facility (EFSF); credit value adjustment (CVA); collateralized debt obligation (CDO).

1 INTRODUCTION

Correlation-sensitive instruments (ie, instruments with embedded exposure to credit correlation) constitute a major and growing share of global financial markets. Chief among them are of course securities, which allocate the default risk of a pool of...
underlying credit assets to different tranches with different seniorities under default. The category includes not only the notorious collateralized synthetic obligations (CSOs), or synthetic collateralized debt obligations (CDOs), based on a portfolio of single-name credit default swaps (CDSs), but also their cash-based cousins, such as collateralized loan obligations (CLOs) – staging a comeback among yield-hungry investors in Europe – and asset-backed securities (ABSs). However, the same principle of mutualization of potential losses and the dependence on the probability of joint defaults inherent in CDOs is present in (superficially much simpler) EFSF bonds, ie, bonds issued since 2011 by the European Financial Stability Facility to finance loan programs for European countries hit particularly hard by the sovereign debt crisis. Ironically, even central clearing counterparties (CCPs), envisaged after the crisis to be the main vehicle of managing systemic risk, share the same underlying feature of being essentially “short credit correlation”. Although the mutualization of losses via the CCP default fund can effectively absorb idiosyncratic and independent risks, the largely regulation-imposed concentration of exposures makes CCPs very sensitive to systemic, or correlated, failures, much like a senior tranche of a CDO. Moreover, the risk inherent in CCPs is wrong-way risk, since losses will tend to hit the default fund precisely when clearing members are under financial stress, and even surviving clearing members might eventually fail as a result of losses to their default funds and other loss-allocation methods.

All this highlights the need for robust models of default correlation, which is a key driver of valuation and risk management of correlation-sensitive instruments and financial vehicles. Indeed, growing awareness of these issues has triggered major changes in regulations, accounting and business practices, all targeted at better management of portfolio concentration and default correlation.

For example, Basel 2.5 and Basel III impose the so-called comprehensive risk measure (CRM) to calculate the capital requirements for structured credit positions, making new single-CSO issuance more capital intensive for banks than it was pre-crisis. Similarly, in its “Review of the credit valuation adjustment risk framework” (Basel Committee on Banking Supervision 2015) the Bank for International Settlements (BIS) urges banks to properly account for any wrong-way risk, ie, dependence between risk exposure and the counterparty’s credit quality. Finally, the international body overseeing clearing arrangements (the BIS Committee on Payment and Market Infrastructures and the International Organization of Securities Commissions (CPMI–IOSCO)) has issued “Principles for financial market infrastructures” (2012), recommending that CCPs “should measure and monitor the correlation between a counterparty’s creditworthiness and the collateral posted” whereby “correlation should not be understood to be limited to linear correlation, but rather to encompass a broad range of co-dependence or co-movement in relevant economic variables.”
The industry standard for such applications, however, is based on the Gaussian copula framework (see Andersen and Sidenius (2004), Hull and White (2004) or any of the classic textbooks: O’Kane (2008), for example). Unfortunately, the Gaussian copula is not very well suited to analyzing the concentration of defaults in time. As argued by Morini (2011), the Gaussian copula allows paradoxical and misleading results such as an inverse relation between correlation and the model probability of loss concentration. Copula methods are also difficult to square with the no-arbitrage pricing methodologies applied for other asset classes, making consistent portfolio credit risk monitoring challenging. As a remedy to the failings of the Gaussian copula, Morini proposes the Marshall and Olkin (1967) fatal shock model, which, although it appears to have some theoretical advantages over the Gaussian copula, is not very useful as a portfolio credit risk model in practice (see, in particular, Andersen and Sidenius (2004) for a discussion of problems associated with calibrating a Marshall–Olkin model).

Against this background, we propose a simple model for default correlation in the spirit of Marshall and Olkin, free from the flaws inherent in copulas. In our model, counterparties’ default times are determined by idiosyncratic and systematic factors, whereby a name defaults if it is struck by either factor for the first time. The novelty of the approach lies in representing systematic factors as increasing sequences of random variables, which allows us to build a rich correlation structure and ensures consistency with pricing models developed for other asset classes. The construction of systematic factors preserves the stopping-time property of defaults that is lost both in copulas and in the Marshall–Olkin model. Thus, credit risk can be modeled in the familiar mathematical framework of martingale methods. Finally, in our model, default correlation – the key parameter affecting the value of any correlation-sensitive instrument – is an output, not an input.

The paper proceeds as follows. In Section 2 we briefly describe the Gaussian copula framework and discuss its main shortcomings as a default correlation model. Section 3 presents our model of default correlation. In Section 4 we discuss model calibration and its application to a wide range of correlation-sensitive instruments, from simple repos, a CDS portfolio with wrong-way risk and EFSF bonds to synthetic CDOs. Section 5 concludes.

2 THE GAUSSIAN COPULA MODEL AND ITS SHORTCOMINGS

Default correlation is typically defined as the risk of two or more credits defaulting jointly, with a probability that is greater than if they were independent. However, as well as a “space dimension”, whereby such a correlation is understood as a concentration of losses in a given portfolio, default correlation has a “time dimension”, where it is viewed as the concentration of losses in time, possibly across different...
portfolios. This distinction can be crucial. For example, a cash provider in a repo transaction does not suffer losses if both the counterparty and the collateral issuer default before the maturity of the repo, but rather sustains losses only if both default within a short period of time and in a particular order (first the counterparty and then, during the realization period needed to liquidate the pledged collateral in an orderly fashion, the issuer). A good model should ideally capture both aspects of correlation consistently. To motivate further discussion, we begin by presenting what seems to be the industry standard in modeling default correlation, namely, the Gaussian copula model, a variant of which is implemented, eg, by Bloomberg as a CDO pricing tool.

In the model each obligor $i$ is assigned a standard normal variable $A_i$ and a time-dependent default threshold $z_i(T)$. It is assumed that default occurs before time $T$ if the variable $A_i$ (which itself is not observable and has no dynamics) finds itself below the threshold $z_i(T)$. Formally,

$$P(\tau_i \leq T) = P(A_i \leq z_i(T)) = PD(T),$$

where $\tau_i$ is the default time of credit $i$, and $PD(T)$ is the $T$-year probability of default, calibrated to an obligor’s CDS curve.

The correlation structure in the model is introduced by assuming that all $A_i$ are driven by a common systematic factor, $Z$, and a name-specific idiosyncratic factor, $Y$:

$$A_i = \rho_i Z + \sqrt{1 - \rho_i^2} Y_i,$$

where $Z$ and $Y_i$ are standard normal with $\text{cov}(Y_i, Y_j) = 0$ and $\text{cov}(Y_i, Z) = 0$. Now the vector $A = [A_1, A_2, \ldots, A_n]$ is multivariate Gaussian with correlation given by

$$\text{corr}(A_i, A_j) = \rho_i \rho_j.$$

Denoting the distribution of $A$ by

$$P(A_1 \leq a_1, \ldots, A_n \leq a_n) \equiv \Phi(a_1, \ldots, a_n),$$

we can define the Gaussian copula as

$$C_{A_1, \ldots, A_n}(u_1, \ldots, u_n) = \Phi(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)) = \mathbb{E}(P(A_1 \leq \Phi^{-1}(u_1), \ldots, A_n \leq \Phi^{-1}(u_n)) | Z).$$

---

1 Note that asset correlation is a different concept from default time correlation $\rho(\tau_i, \tau_j)$ and default indicator correlation $\rho(I_i, I_j)$. In general, conditional on default probabilities, default correlation is an increasing function of asset correlation (Hanson et al 2008).
Since, for a given $Z$, the $A_i$ are independent, the copula can be conveniently re-expressed as

$$C_{A_1, \ldots, A_n} = \mathbb{E} \left( \prod_{i=1}^{n} \mathbb{P}(A_i < \Phi^{-1}(u_i)) \mid Z \right)$$

$$= \mathbb{E} \left( \prod_{i=1}^{n} \mathbb{P}(\rho_i Z + \sqrt{1 - \rho_i^2} Y_i \leq \Phi^{-1}(u_i)) \mid Z \right)$$

$$= \mathbb{E} \left( \prod_{i=1}^{n} \Phi \left( \frac{\Phi^{-1}(u_i) - \rho_i Z}{\sqrt{1 - \rho_i^2} Y_i} \right) \right)$$

$$= \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Phi \left( \frac{\Phi^{-1}(u_i) - \rho_i Z}{\sqrt{1 - \rho_i^2} Y_i} \right) \phi(Z) \, dZ,$$

where $\phi(Z)$ is the standard normal density. The probability of joint default is now given by

$$\mathbb{P}(\tau_1 \leq T_1, \ldots, \tau_n \leq T_n) = \int_{-\infty}^{\infty} \prod_{i=1}^{n} \Phi \left( \frac{\Phi^{-1}(PD(T_i)) - \rho_i Z}{\sqrt{1 - \rho_i^2} Y_i} \right) \phi(Z) \, dZ. \quad (2.3)$$

Although (2.3) is very convenient numerically (it involves only one-dimensional integration), this simplicity comes at the cost of reducing the entire dependence structure of a portfolio of $n$ credits to linear correlations determined by $\frac{1}{2} n (n - 1)$ parameters. In practice, to avoid technical difficulties, popular implementations of the model consist in setting all $\rho_i$ equal to one parameter, $\rho$. For example, the market standard (the so-called base correlation model) for valuing CDOs assumes that each attachment/detachment point has its own input level of $\rho$, so that a mezzanine tranche $[X, Y]$ is expressed as the difference between two equity tranches $[0, Y]$ and $[0, X]$, each valued with different correlation parameters.\(^2\)

The Gaussian copula model has three major shortcomings as a tool for modeling correlated defaults (see Morini (2011) and Gaątarek and Jabłecki (2015) for an extensive critique of the Gaussian copula approach).

1. The model is incapable of producing a concentration of defaults in time if obligors in the portfolio have different conditional default probabilities. As a result, for some parameterizations, even perfect asset correlation does not produce default clustering.

\(^2\) Note that this assumption implies a major inconsistency, in that different equity tranches assign different correlations to the same underlying reference portfolios.
(2) The relationship between the probability of joint defaults, implied by the model, and asset correlation is unstable, i.e., the probability of joint defaults can be either an increasing function or a decreasing function of asset correlation depending on the time frame under consideration and the obligors’ PDs.

(3) Related to the previous point, the probability of defaults occurring consecutively and close together in time, as, for example, in a repo double default scenario, can, for given conditional hazard rates, be a nonmonotonic hump-shaped function of the asset correlation. The problem with such a nonmonotonic pattern is not only that it is difficult to understand from an economic point of view, but also that it makes it generally impossible to know a priori whether, with a given set of PDs, an increase in correlation will increase or decrease central bank risk exposure.

These problems are generally known to practitioners and can sometimes be circumvented by making straightforward tweaks to the model. However, the Gaussian copula does not allow for an intuitive relationship between the input parameters (such as PDs and asset correlations) and the model output, which can be a major source of model risk.

3 A SIMPLE FACTOR MODEL OF JOINT DEFAULTS

In this section we propose our own model for correlated defaults that is free from the inherent shortcomings of the Gaussian copula approach. To facilitate exposition, we first introduce the model in the simplest single-factor and constant-hazard-rate setting before offering a more general perspective.3

3.1 Single-factor case

Consider $d$ obligors with default times $\tau_1, \ldots, \tau_d$. Assume for now (we relax this assumption in Section 3.2) that default times are exponentially distributed with parameters $\lambda_1, \ldots, \lambda_d$, which admit natural interpretation as hazard rates or conditional default probabilities. As before, we introduce dependence between default times by stating that each default can result from the materialization of either an idiosyncratic factor or a systematic factor, whichever hits sooner. Being hit by either factor can be interpreted mathematically as the first jump in a specific Poisson process. Hence, for each obligor $i$ the time until the arrival of the idiosyncratic factor is represented simply by an exponential variable, $Y_i$, with parameter $\lambda_i^{\text{idio}}$. Where we differ from previous approaches is that, unlike in (2.2), where the systematic factor was a single random

---

3 The model was originally introduced in Gątarek and Jabłecki (2013, 2015), and the summary below draws on both accounts.
variable, we think of a systematic factor as an increasing sequence of exponential variables \( Z_1 \leq \cdots \leq Z_d \) with parameters \( \lambda_1^{\text{sys}}, \ldots, \lambda_d^{\text{sys}} \). This should be intuitive, as the most natural interpretation of dependence for random variables expressing time is an ordering relation. Under such assumptions, individual obligors’ default times can be represented as

\[
\tau_i = \min\{Y_i, Z_i\},
\]

(3.1)

where \( Y_1, \ldots, Y_d \) and \( Z_1, \ldots, Z_d \) are independent exponential variables. Obviously, default times of all obligors \( \tau_i \) are also exponentially distributed with parameters \( \lambda_i = \lambda_i^{\text{idio}} + \lambda_i^{\text{sys}} \) and survival probabilities

\[
P(\tau_i > T) = e^{-\lambda_i T}.
\]

(3.2)

Consider now two useful properties of the proposed model.

**Remark 3.1** Idiosyncratic defaults tend to be more frequent than systematic defaults.

Denote the first default of an idiosyncratic type by \( Y_{\text{first}} = \min\{Y_i : 1 \leq i \leq d\} \) and the first default of a systematic type by \( Z_1 \). Then

\[
P(Y_{\text{first}} \geq t) = P(\min\{Y_i : 1 \leq i \leq d\} \geq t) = \exp \left( -t \sum_{i=1}^{d} \lambda_i^{\text{idio}} \right).
\]

(3.3)

Since systematic defaults are, by definition, ordered, we know which obligor defaults first (although we do not know exactly when):

\[
P(Z_1 \geq t) = \exp(-t \lambda_1^{\text{sys}}).
\]

(3.4)

Hence,

\[
\frac{P(t \leq Z_t < t + dt \mid Z_1 > t)}{P(t \leq Y_{\text{first}} < t + dt \mid Y_{\text{first}} > t)} = \frac{\lambda_1^{\text{sys}}}{\sum_{i=1}^{d} \lambda_i^{\text{idio}}}.
\]

(3.5)

and, under normal conditions, the chance that the first default is idiosyncratic should be considerably greater.

**Remark 3.2** The definition of a systematic factor as an increasing sequence of random variables allows us to capture the phenomenon of default clustering.

---

4 We show below that defining a systematic factor as an increasing family of random variables rather than a single random variable allows us to preserve the useful formal property that default times are also stopping times. Such a redefinition also formalizes the important empirical idea that a systematic factor need not cause the default of the whole economy at once. It is much more natural to expect that some credits will default sooner and some later, depending on their sensitivity to the given systematic factor.
In fact, in our model only systematic defaults can be multiple defaults. To see this, define the following point processes, counting the defaults triggered by the respective factors \( N(t) = \text{card}(\tau_i < t) \), \( N^{\text{sys}}(t) = \text{card}(Z_i < t) \) and \( N_j(t) = \text{card}(i > j : Y_i < t) \). Using the property that \( Z_i \) are ordered, we easily get

\[
P(N^{\text{sys}}(t) = j) = P(Z_j < t < Z_{j+1}) = (1 - e^{-t\lambda^{\text{sys}}_j}) - (1 - e^{-t\lambda^{\text{sys}}_{j+1}}) = e^{-t\lambda^{\text{sys}}_{j+1}} - e^{-t\lambda^{\text{sys}}_j}.
\]

Consequently,

\[
P(N(t) = m) = \sum_{j=0}^{m} P(N^{\text{sys}}(t) = j)P(N_j(t) = m - j) = \sum_{j=0}^{m} (e^{-t\lambda^{\text{sys}}_{j+1}} - e^{-t\lambda^{\text{sys}}_j})P(N_j(t) = m - j).
\]

If we assume that defaults of individual obligors can repeat themselves,\(^5\) then the point process \( N_j(t) \) is a Poisson process with intensity \( \sum_{i=j+1}^{d} \lambda_i^{\text{idio}} \). Hence,

\[
P(N_j(t) = m) \approx \frac{1}{m!} \left( t \sum_{i=j+1}^{d} \lambda_i^{\text{idio}} \right)^m \exp \left( -t \sum_{i=j+1}^{d} \lambda_i^{\text{idio}} \right).
\]

Figure 1 shows how the ordering of random variables comprising a systematic factor affects the aggregate default distribution. The same set of 100 hazard rates (ranging from 10% to 0.6%) and (3.6)–(3.8) are used to produce default distributions, assuming that default times are idiosyncratic in one case and that they are ordered in the other. The ordering of default times clearly increases the chance of having multiple defaults within a given time horizon.

### 3.2 Multi-factor extension

We now show how to generalize the simple model proposed above to cater for a greater number of factors, a larger class of factor distributions and stochastic hazard rates. For modeling purposes, we assume that all processes and variables are defined on a filtered probability space \((\Omega, \mathcal{F}, P)\), under the usual conditions, with \((\mathcal{F}_t)_{t \geq 0}\) modeling the information flow and \( P \) being the risk-neutral (martingale) measure relative to which

---

\(^5\) Since, in practice, defaults of individual obligors cannot repeat themselves, which is equivalent to randomly drawing defaulting names without replacement, this assumption slightly overstates the total number of defaults. A more rigorous approach to estimating such probabilities is based on the Bernoulli triangle as suggested originally by Hull and White (2004).
FIGURE 1 Default distribution for a portfolio of 100 credits assuming idiosyncratic and systematic defaults.

The same set of 100 hazard rates is used to generate both plots, but the shocks are treated as idiosyncratic in one case, and as systematic in the other.

all security prices discounted by the risk-free interest rate are martingales. A stopping, or default, time with respect to $\mathcal{F}_t$ is a random variable, $\tau$, such that $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \geq 0$. We say that a nonnegative, $\mathcal{F}_t$-predictable process $\lambda(t)$ is the hazard rate (equivalently, default intensity) of $\tau$ if $1_{\{\tau > t\}} + \int_0^t \lambda(s) \, ds$ is a martingale (see Brémaud (1981) or Schönbucher (2003) for details).

We begin with a general definition of a factor.

**Definition 3.3** By a systematic factor we mean a pair $(Z_i)_{i=1}^d, \Phi)$, where $(Z_i)_{i=1}^d$ is a family of positive random variables with given distributions, and $\Phi$ is a permutation of the set $\{1, 2, \ldots, d\}$ such that the sequence $Z_{\Phi(i)}$ is increasing, ie, $Z_{\Phi(i)} \leq Z_{\Phi(i+1)}$ for $i = 1, \ldots, d$.

Definition 3.3 extends the concept of a factor as an ordered family of random variables to a multifactor setting where each factor is characterized by the order in which it triggers the default of a sequence of risky names. The ordering of defaults – reflecting different sensitivities to different factors – is modeled by associating each factor (ie, ordering) with a permutation of the set of names $\{1, 2, \ldots, d\}$. Since each factor is given by a different permutation of $\{1, \ldots, d\}$, there are $d!$ possible systematic factors for a given set of obligors. The reason why the Marshall–Olkin “fatal shocks” approach has important practical limitations has been explained above. Here we stress the mathematical advantage of our definition, which is that it allows different obligors to have different correlations with the systematic factors (ie, different factor loadings)
while preserving the “stopping time” character of credit events. The latter, in turn, ensures the mathematical flexibility of martingale pricing and consistency with models developed for other asset classes, thus in principle allowing the central bank to use a consistent model for both policy and investment operations.

The construction of systematic factors in a multifactor setting is analogous to that presented in Section 4. Specifically, let \( Z = \{ \tilde{Z}_i : i = 1, \ldots, d \} \) be a set of \( \mathcal{F}_t \) measurable stopping times with respective conditionally independent hazard rates \( \lambda(t) \), ie, such that

\[
1_{\{\min_{\xi \in V} \xi > t\}} + \int_0^t 1_{\{\min_{\xi \in V} \xi > s\}} \sum_{\xi \in V} \lambda_{\xi}(s) \, ds
\]

(3.9)
is a martingale for any \( V \subseteq \mathbb{Z} \). Let \( \Phi(\cdot) \) be a permutation of the set \( \{1, \ldots, d\} \). Then, the family \( Z_i = \min\{\tilde{Z}_j : \Phi^{-1}(i) \leq j \leq d\} \) is a systematic factor. The sequence \( Z_{\Phi(i)} = \min\{\tilde{Z}_i : i \leq j \leq d\} \) is increasing, and, by assumption (3.9),

\[
N_i(t) + \int_0^t N_i(s) \sum_{j=\Phi^{-1}(i)}^d \lambda_j(s) \, ds
\]

(3.10)
is a martingale and \( \sum_{j=\Phi^{-1}(i)}^d \lambda_j(t) \) is the hazard rate of \( Z_i \). To construct another factor, we start again with a sequence of \( d \) positive random variables, choose a new permutation \( \Psi \) of the set \( \{1, \ldots, d\} \) such that \( \Psi \neq \Phi \) and proceed with the construction exactly as before. Once again, note that such a representation ensures, in a clear and mathematically tractable way, that each of the \( \{1, \ldots, d\} \) names in the economy can have a different dependence on one of the \( d! \) systematic factors. This allows an almost arbitrarily rich correlation structure, certainly beyond that implied by market data before the crisis.

We can now present the definition of correlated default times.

**Definition 3.4** Let \( U = \{Y_i, Z^1_i, \ldots, Z^N_i : 1 \leq i \leq d\} \) be the set of default times of all risk factors, both idiosyncratic, \( Y_i \), and systematic \( Z^1_i, \ldots, Z^N_i \), conditionally independent of their hazard rates. We write

\[
U(i) = \{Y_i, Z^j_k : \Phi^{-1}_j(i) \leq k < d, \ 1 \leq j \leq N\} \subseteq U,
\]

Note that this is not true in the Marshall–Olkin fatal-shock model, ie, all factors have, by definition, the same sensitivity to the systematic factor. Our construction is the only way to ensure the stopping-time property while allowing different systematic factor loadings. To see this, suppose by contradiction that we represent default times by \( \tau_i = \min\{Y_i, k_i Z\} \), where the variables \( k_i \) represent the sensitivity of name \( i \) to factor \( Z \), which is here a standard exponential variable. Then, if \( k_i < 1 \), we have that \( t/k_i > t \) and \( \{k_i Z \leq t\} = \{Z \leq t/k_i\} \notin \mathcal{F}_t \). Hence, \( k_i Z \) is not a stopping time, and by implication \( 1/k_i \) is not a hazard rate.
where $\Phi_j(\cdot)$ are the permutations ordering $(Z_{i,j}^j)_{i=1}^d$, $j = 1, \ldots, N$ and we define dependent default times as

$$\tau_i = \min\{Y_i, Z_i^1, \ldots, Z_i^N\} = \min_{\xi \in U(i)} \xi. \tag{3.11}$$

The assumptions now guarantee that for all $t < T$ conditional survival probabilities under the natural filtration $\mathcal{F}_t$ are given by

$$\mathbb{P}(\xi > T \mid \mathcal{F}_t) = \mathbb{E}\left\{ \exp\left( - \int_t^T \lambda_\xi(s) \, ds \right) \mid \mathcal{F}_t \right\} 1_{\{\xi > t\}} \tag{3.12}$$

and

$$\mathbb{P}(\tau_i > T \mid \mathcal{F}_t) = \mathbb{E}\left\{ \exp\left( - \int_t^T \sum_{\xi \in U(i)} \lambda_\xi(s) \, ds \right) \mid \mathcal{F}_t \right\} 1_{\{\tau_i > t\}}. \tag{3.13}$$

### 4 EXAMPLES AND APPLICATIONS

Calibration of the proposed model entails distributing the market-implied (or otherwise obtained) hazard rates of individual obligors across the idiosyncratic and systematic shock components. The first step in the process is the construction of a systematic factor, which can, for example, be done in the following way. Start by ordering obligors according to their overall exposure to the systematic factor in question. Suppose, for example, that credit “1” is the most sensitive to the systematic factor, credit “2” is less so but still highly sensitive, and so on, while credit $d$ is the least exposed. Hence, the systematic factor should first trigger the default of name “1”, then “2”, etc, before ultimately hitting $d$. To reflect this, assign to each name $i$ a Poisson process $\tilde{Z}_i$, with intensity $\lambda_i^{sys}$, whose arrival triggers the default not only of credit $i$, but also, due to the ordering relation, of all the more systematically risky names $i_1, i_2, \ldots, i_d$ (we assume that Poisson processes $\tilde{Z}_i$ are independent). Thus, the systematic intensity of each obligor $i$ will be the sum of its own intensity $\lambda_i^{sys}$ and the intensities of the Poisson processes triggering defaults of more senior names, ie, $\sum_{j=i}^d \lambda_j^{sys}$. This can be formalized by setting

$$Z_i = \min\{\tilde{Z}_i : i \geq j\}, \tag{4.1}$$

where $Z_i$ is the Poisson process representing the total systematic exposure of obligor $i$. Note that $Z_i \leq Z_{i+1}$ for $i = 1, \ldots, d - 1$, so indeed the family $Z_1, \ldots, Z_d$ is a systematic factor. Since each obligor is also affected by an idiosyncratic shock $Y_i$ with intensity $\lambda_i^{idio}$, default times $\tau_i$ are exponentially distributed with parameters...
\[ \lambda_i = \lambda_i^{\text{idio}} + \sum_{j=i}^{d} \lambda_j^{\text{sys}} \] and survival probabilities

\[ P(\tau_i > T) = P(\min\{Y_i, \min\{Z_i, \tilde{Z}_{i+1}, \ldots, \tilde{Z}_d\}\} > T) \]
\[ = P(\min\{Y_i, \tilde{Z}_i, \tilde{Z}_{i+1}, \ldots, \tilde{Z}_d\} > T) \]
\[ = P(Y_i > T)P(\tilde{Z}_i > T)P(\tilde{Z}_{i+1} > T) \cdots P(\tilde{Z}_d > T) \]
\[ = e^{-\lambda_i T}. \] (4.2)

The second step of the calibration process consists in allocating the total hazard rate of each obligor (obtained from, eg, rating agencies, market spreads or in-house credit models) to the idiosyncratic and the systematic component. In what follows we show various ways in which this can be done and apply the model in four different contexts: estimating residual credit risk on a repo portfolio; estimating wrong-way risk in a CDS portfolio; pricing an EFSF bond; and, finally, pricing CDO tranches.

4.1 Joint defaults in a CCP repo portfolio

Repurchase agreements (repos) constitute a major share of global money markets – up to €15 trillion in size, according to the International Capital Market Association, with a growing share of contracts cleared through CCPs. In a typical repo, one party (the cash taker) sells an asset to another party (cash provider) with a promise to repurchase that asset at some future date. Although a repo, like any other financial instrument, is subject to counterparty credit risk, market risk, liquidity risk and operational risk, all these risks can be substantially reduced by a selection of risk management tools, including haircuts on underlying securities, mark-to-market valuation, as well as counterparty and collateral limits (see, for example, Choudhry (2002) for an extensive overview). If the risk mitigation tools are properly deployed (which is what we shall assume here), the cash provider is effectively left only with a small residual credit risk that can materialize in so-called double default events when the counterparty and the collateral issuer both default within a short period of time. As stressed above, the time frame (and the order of defaults) is crucial, as in most cases it would not be advisable to immediately sell the collateral captured after counterparty default, because such a sale could adversely affect market prices (which may already be affected by the counterparty default event itself), leading to potential mark-to-market losses – especially for a CCP handling large and concentrated collateral portfolios. We now demonstrate how our model can be useful in accurately accounting for residual credit risk in a repo book run by a large clearing house.\(^7\) Although the example is stylized it demonstrates the key advantages of the proposed modeling approach.

\(^7\) See also Gątarek and Jabłecki (2015) for a more detailed discussion of repo risk management.
### TABLE 1 Repo transactions of EU and US banks.

<table>
<thead>
<tr>
<th>No.</th>
<th>CCP counterparty</th>
<th>5Y CDS (bps)</th>
<th>Toy CCP portfolio share (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Barclays</td>
<td>136.18</td>
<td>11.9</td>
</tr>
<tr>
<td>2</td>
<td>BNP Paribas</td>
<td>112.655</td>
<td>9.6</td>
</tr>
<tr>
<td>3</td>
<td>BoA</td>
<td>141.688</td>
<td>7.8</td>
</tr>
<tr>
<td>4</td>
<td>RBS</td>
<td>132.792</td>
<td>7.4</td>
</tr>
<tr>
<td>5</td>
<td>Credit Suisse</td>
<td>171.976</td>
<td>6.8</td>
</tr>
<tr>
<td>6</td>
<td>GS</td>
<td>150.414</td>
<td>6.8</td>
</tr>
<tr>
<td>7</td>
<td>Credit Agricole</td>
<td>113.383</td>
<td>6.1</td>
</tr>
<tr>
<td>8</td>
<td>Deutsche Bank</td>
<td>221.989</td>
<td>6.0</td>
</tr>
<tr>
<td>9</td>
<td>Citi</td>
<td>142.964</td>
<td>6.0</td>
</tr>
<tr>
<td>10</td>
<td>Société Générale</td>
<td>112.578</td>
<td>5.3</td>
</tr>
<tr>
<td>11</td>
<td>HSBC</td>
<td>135.871</td>
<td>4.5</td>
</tr>
<tr>
<td>12</td>
<td>UBS</td>
<td>90.713</td>
<td>3.9</td>
</tr>
<tr>
<td>13</td>
<td>MS</td>
<td>151.945</td>
<td>3.8</td>
</tr>
<tr>
<td>14</td>
<td>Unicredit</td>
<td>232.614</td>
<td>3.3</td>
</tr>
<tr>
<td>15</td>
<td>Santander</td>
<td>183.179</td>
<td>3.2</td>
</tr>
<tr>
<td>16</td>
<td>BBVA</td>
<td>182.466</td>
<td>2.0</td>
</tr>
<tr>
<td>17</td>
<td>Danske</td>
<td>92.232</td>
<td>1.7</td>
</tr>
<tr>
<td>18</td>
<td>Commerz</td>
<td>156.915</td>
<td>1.4</td>
</tr>
<tr>
<td>19</td>
<td>Lloyds</td>
<td>101.818</td>
<td>1.3</td>
</tr>
<tr>
<td>20</td>
<td>Wells Fargo</td>
<td>76.077</td>
<td>1.1</td>
</tr>
</tbody>
</table>

For illustration purposes, suppose the CCP repo book is made up of exposures toward the most active players in the global repo market as reported in the 2015 ICMA repo market survey (Table 1). Assume also that the credit quality of collateral posted reflects averages reported by tri-party agents in the survey, ie, AAA (30% of total collateral posted), AA (30%), A (10%), BBB (13%), BBB— and below (17%). Finally, assume that the liquidation period is uniform and set by the CCP at seven days, corresponding to that used by major European CCPs. Now the calculation of portfolio residual credit risk amounts to simulating the following:

$$\mathbb{P}(\{\tau_{C_1} \leq 3M \wedge |\tau_{C_1} - \tau_{A_1}| \leq 7D\} \vee \cdots \vee \{\tau_{C_{20}} \leq 3M \wedge |\tau_{C_{20}} - \tau_{A_{20}}| \leq 7D\})$$

(4.3)

where $\tau_{C_i}$ and $\tau_{A_i}$ denote the default times of counterparty $C_i$ and collateral asset $A_i$, respectively. Such probability will be driven by the level of portfolio correlation: in other words, by the degree to which portfolio defaults are determined by systematic and idiosyncratic factors. For stress testing it can be convenient to capture portfolio correlation using a single parameter, $\rho \in [0, 1]$, with $\rho = 0$ implying totally independent defaults and $\rho = 1$ completely comonotonic defaults. The correlation parameter...
\( \rho \) determines the individual systematic exposure of each name in the portfolio, taking into account the cascading nature of the systematic shock, through the following identities:

\[
\begin{align*}
\lambda_{20}^{\text{sys}} &:= \rho \lambda_{20}, \\
\lambda_{19}^{\text{sys}} &:= \rho \lambda_{19} - \lambda_{20}^{\text{sys}} = \rho (\lambda_{19} - \lambda_{20}), \\
\lambda_{18}^{\text{sys}} &:= \rho \lambda_{18} - \lambda_{19}^{\text{sys}} - \lambda_{20}^{\text{sys}} = \rho (\lambda_{18} - \lambda_{19}), \\
\vdots \\
\lambda_{1}^{\text{sys}} &:= \rho \lambda_{1} - \lambda_{2}^{\text{sys}} - \cdots - \lambda_{20}^{\text{sys}} = \rho (\lambda_{1} - \lambda_{2}),
\end{align*}
\]

(4.4)

where \( \lambda_1, \ldots, \lambda_{20} \) are hazard rates of CCP counterparties, sorted in ascending order of systematic riskiness. Hence, counterparty \( C_{20} \) has the lowest systematic riskiness and systematic shock intensity given simply by \( \rho \lambda_{20} \).\(^8\) Counterparty \( C_{19} \) is slightly more risky with a total intensity comprising both the individual systematic sensitivity \( \lambda_{19}^{\text{sys}} \) and the default intensity of the less risky obligor \( \lambda_{20}^{\text{sys}} \). Finally, the most systematically risky obligor has total intensity equal to the sum of its individual systematic sensitivity and the intensities of all better credits, which reduces to \( \rho (\lambda_1 - \lambda_2) \). On top of that, each obligor is also hit by an independent idiosyncratic shock whose intensity is determined by subtracting the total systematic intensity from that name’s hazard rate, ie,

\[
\lambda_{i}^{\text{idio}} := \lambda_{i} - \sum_{j=1}^{d} \lambda_{j}^{\text{sys}} = (1 - \rho) \lambda_{i}.
\]

(4.5)

Although the true \( \rho \) is obviously not observable, we determine a proxy in the following way. First, we run principal components analysis (PCA) on CDS spreads of the twenty most active US and EU banks in the repo market (Table 1). We then associate \( \rho \) with the percentage of variance explained by the PCA factor with the most consistently positive loadings (loadings on the systematic factor, as opposed to idiosyncratic factors, should generally have the same sign). In case of our “toy” CCP portfolio \( \rho = 0.80 \), which puts the probability of joint defaults at about 30 basis points (bps). To check how sensitive our estimates are to different levels of \( \rho \), we can easily rerun the analysis, changing the decomposition of hazard rates in the portfolio as per (4.4) each time. Figure 2 shows that – in line with our intuition – the probability of joint defaults in our toy CCP portfolio is an increasing function of the level of portfolio correlation.

\(^8\)To simplify presentation we omit time subscripts and assume constant and deterministic hazard rates. In practice, a nontrivial term structure of hazard rates should be calibrated.
4.2 Wrong-way risk in a CDS portfolio

As another example we show how our default correlation model can be used to account for wrong-way risk in calculating the credit value adjustment (CVA) for a CDS portfolio, such as the one recently purchased by Citi from Deutsche Bank (Devasabai 2015). As before, the exposition is highly stylized, but it demonstrates a pattern that will also hold in more realistic settings. Recall first that CVA is an adjustment to the fair value (or price) of derivatives to account for counterparty credit risk or, equivalently, the cost of hedging the risk of counterparty default. The tacit assumption of many CVA models is that counterparties’ default times are independent of all economic factors driving the value of a derivatives portfolio. Although this assumption greatly simplifies calculations, it can produce misleading results when applied in the context of credit derivatives, where the same systematic factors are likely to affect position value (ie, essentially creditworthiness of the reference names) and counterparty creditworthiness. A good CVA model for credit derivatives must therefore account for the so-called wrong-way risk, ie, situations when default risk increases with the position value. This, in turn, requires linking the value of the underlying contract with the default of the counterparty in a coherent way – something that can naturally be achieved within the framework we propose. The first step entails deriving counterparty-risk-adjusted pricing formulas for credit default swaps. Consider a CDS contract referencing name $i$ and sold by name $j$. The contract has cashflows $p \delta_n \mathbf{1}_{\{X_i > t_n\}} \mathbf{1}_{\{X_j > t_n\}}$ made on each of the payment dates $t_1, t_2, \ldots, t_N = T$, where $\delta_n$ is the day count fraction for the period $[t_{n-1}, t_n]$, and $\mathbf{1}_{\{X_i > t_n\}} \mathbf{1}_{\{X_j > t_n\}}$ reflects the fact that payments stop when either the counterparty or the reference entity defaults. The contingent protection leg is a single payment of $1 - R$ times notional in the event
of \( i \)'s default and \( j \)'s survival, \( R \) being the recovery rate on the defaulted asset. So the contract value at time \( t \) is (assuming unit notional):

\[
\text{CDS}(t, T) = (1 - R) \mathbb{E} \left( \int_t^T D(s) \mathbf{1}_{\{X_j > s\}} d \mathbf{1}_{\{X_i > s\}} \bigg| \mathcal{F}_t \right) \\
- \mathbb{E} \left( \sum_{n=1}^N p \delta_n D(t_n) \mathbf{1}_{\{X_i > t_n\}} \mathbf{1}_{\{X_j > t_n\}} \bigg| \mathcal{F}_t \right).
\]

The fair value of the spread is found by setting this contract value to zero at inception, \( \text{CDS}(0, T) = 0 \), which ultimately produces the following formula (see Gątarek and Jabłecki 2013):

\[
\text{CDS}(t, T) = (1 - R) \mathbb{E} \left( \int_t^T G(s) \sum_{\tau \in U(i) \setminus U(j)} \lambda_\tau(s) \, ds \bigg| \mathcal{F}_t \right) D(t) \mathbf{1}_{\{X_i > t\}} \mathbf{1}_{\{X_j > t\}} \\
- \mathbb{E} \left( \sum_{n=1}^N p \delta_n \exp \left( - \int_0^{t_n} r(s) + \sum_{\tau \in U(i) \cup U(j)} \lambda_\tau(s) \, ds \right) \bigg| \mathcal{F}_t \right) \\
\times D(t) \mathbf{1}_{\{X_i > t\}} \mathbf{1}_{\{X_j > t\}},
\]

where

\[
G(s) = \exp \left( - \int_t^s r(u) + \sum_{\tau \in U(i) \cup U(j)} \lambda_\tau(u) \, du \right).
\]
### Table 2: Composition of the CDS portfolio

<table>
<thead>
<tr>
<th>(a) Companies</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Hazard rate (%)</strong></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>4.07</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3.58</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2.43</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2.18</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1.93</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(b) Counterparties</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
<td><strong>Hazard rate (%)</strong></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1.43</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1.37</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>

To illustrate how default correlation (i.e., dependence on systematic factors) affects counterparty risk and alters the valuation of a CDS that counterparty “A” sells as protection against the default of a more risky issuer, “B”. Assume a deterministic recovery rate on B’s debt $R$, a deterministic interest rate $r$ and a one-factor version of the dependence model introduced in Definition 3.4, where the systematic factor has default time $\tau_S$ and hazard rate $\lambda_S$. The two names have default times $\tau_A$, $\tau_B$, respectively, with deterministic hazard rates $\lambda_A$, $\lambda_B$ such that $\tau_A = \min\{\tau_1, \tau_S\}$ and $\tau_B = \min\{\tau_2, \tau_S\}$. Obviously, then $\lambda_A = \lambda_1 + \lambda_S$, $\lambda_B = \lambda_2 + \lambda_S$ and $\min\{\tau_A, \tau_B\} = \min\{\tau_1, \tau_2, \tau_S\}$. The spread leg $FL$ of a CDS with counterparty risk equals

$$FL = -p \sum_{k=1}^N \delta_k \exp(-(r + \lambda_1 + \lambda_2 + \lambda_S) t_k) \approx \int_0^T p \exp(-(r + \lambda_1 + \lambda_2 + \lambda_S) t) \, dt,$$

where $p$ is the CDS spread, $t_1 < \cdots < t_n = T$ are fee payment days and $\delta_k = t_k - t_{k-1}$. Similarly, the protection leg $PL$ is given by

$$PL = (1 - R) \int_0^T \lambda_2 \exp(-(r + \lambda_S + \lambda_1 + \lambda_2) t) \, dt.$$

Define default correlation as $\rho = \lambda_S / (\lambda_S + \lambda_1)$ so that if $\rho = 0$ ($\lambda_S = 0$), obligors are independent and if $\rho = 1$ ($\lambda_1 = 0$), obligors’ defaults are comonotonic. Since
Figure 4: Single-name CDS portfolio value as a function of the systematic dependence parameter $\rho$.

\[ \lambda_2 = \lambda_B - \lambda_A \rho, \]  
\[ (4.7) \]  
\[ (4.8) \]

imply that the fair value of the CDS spread can be approximated by

\[ p \approx (1 - R)(\lambda_B - \lambda_A \rho). \]  
\[ (4.9) \]

Hence, the CDS spread is a decreasing function of the default correlation, as shown in Figure 3.

A natural extension of the preceding example involves a portfolio of default swaps. The major problem here is accounting for correlation effects within the entire portfolio. To this end we need to construct a systematic factor, following the logic outlined above for a repo portfolio (as before, we only introduce one factor, but extensions are straightforward). To fix ideas, consider a single-name CDS portfolio consisting of five contracts (with unit notional) written by five counterparties. Concretely, assume the portfolio is made up of counterparty-company positions $N_i, N_{i-5}, i = 5, \ldots, 10$, with deterministic (flat) hazard rates of the respective names given in Table 2. Suppose the maturity of all positions is five years, the recovery rate is 40% and the interest rate equals 2%. Figure 4 shows how the value of our CDS portfolio changes as the systematic exposure, captured by $\rho$, grows from 0% to 100%.

### 4.3 Pricing EFSF bond issuance

As a prelude to pricing an actual CDO, we consider the quasi-CDO bonds issued by the European Financial Stability Facility (EFSF).\(^9\) The EFSF is a company founded

---

\(^9\) We are happy to acknowledge Massimo Morini’s inspiration for applying our model to pricing EFSF issues. Morini was the first to propose pricing the EFSF issue covering a loan to Ireland using
TABLE 3  EFSF shareholder contribution.

<table>
<thead>
<tr>
<th>Member state</th>
<th>S&amp;P rating</th>
<th>Contribution key (%)</th>
<th>Guarantee commitment (€bn)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Austria</td>
<td>AA+</td>
<td>2.9869</td>
<td>21.639</td>
</tr>
<tr>
<td>Belgium</td>
<td>AA</td>
<td>3.7313</td>
<td>27.031</td>
</tr>
<tr>
<td>Cyprus</td>
<td>BB−</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Estonia</td>
<td>AA−</td>
<td>0.2754</td>
<td>1.994</td>
</tr>
<tr>
<td>Finland</td>
<td>AA+</td>
<td>1.9289</td>
<td>13.974</td>
</tr>
<tr>
<td>France</td>
<td>AA</td>
<td>21.8762</td>
<td>158.487</td>
</tr>
<tr>
<td>Germany</td>
<td>AAA</td>
<td>29.1309</td>
<td>211.045</td>
</tr>
<tr>
<td>Greece</td>
<td>B−</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ireland</td>
<td>A+</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Italy</td>
<td>BBB−</td>
<td>19.2233</td>
<td>139.267</td>
</tr>
<tr>
<td>Luxembourg</td>
<td>AAA</td>
<td>0.2687</td>
<td>1.946</td>
</tr>
<tr>
<td>Malta</td>
<td>BBB+</td>
<td>0.0972</td>
<td>0.704</td>
</tr>
<tr>
<td>Netherlands</td>
<td>AAA</td>
<td>6.135</td>
<td>44.446</td>
</tr>
<tr>
<td>Portugal</td>
<td>BB</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Slovakia</td>
<td>A+</td>
<td>1.0666</td>
<td>7.727</td>
</tr>
<tr>
<td>Slovenia</td>
<td>A−</td>
<td>0.5058</td>
<td>3.664</td>
</tr>
<tr>
<td>Spain</td>
<td>BBB+</td>
<td>12.7739</td>
<td>92.543</td>
</tr>
</tbody>
</table>

Cyprus, Greece, Ireland and Portugal have received assistance from the EFSF and their guarantee contributions are amended to zero.

in 2010 to “preserve financial stability of Europe’s monetary union by providing temporary financial assistance to euro area Member States.” (This assistance was provided to Cyprus, Greece, Ireland and Portugal, to the overall tune of €174.6 billion (€130.9 billion of which went to Greece) at the end of February 2016.) The EFSF finances its lending operations by issuing bonds guaranteed jointly by all euro area member states in accordance with their share in the paid-up capital of the European Central Bank (Table 3). The guarantees are “irrevocable and unconditional” and currently gross up to ca. 160.4% over-collateralization, covering both principal and interest on EFSF bonds. Since 2011, funds raised through various issues have not been attributed to a particular country. Instead, all the funds are pooled and then distributed to recipient countries. In the case of a missed payment by a borrower, the EFSF is to ensure that each guarantor remits its share of the shortfall to the EFSF. Hence, pricing EFSF issues ultimately relies on modeling the joint default pattern of a pool of sovereign obligors and its interaction with the credit enhancement mechanisms. This

the Marshall–Olkin model in a presentation called “Pricing EFSF”. Our approach can be seen as a more refined version of his idea, with the modeling of a systematic factor as a chain of dependencies rather than a single random variable.
boils down to estimating the probability of recovering at least 100% of the loan out of a pool of claims amounting to 260.4%. Thus, despite having no tranche structure, EFSF bonds resemble CDOs, and in any case fall into the category of default-correlation-sensitive instruments. Little wonder, then, that the market standard for pricing EFSF issues was the Gaussian copula (this, however, relied on a purely arbitrary notion of “asset correlation” between euro area sovereigns) (see, for example, Moody’s 2010).

In contrast, our approach starts from a qualitative analysis of sovereign risk, which then determines the breakdown of market-implied hazard rates of each country into idiosyncratic and systematic components, making default correlation the output, not the input, of the model.

The starting point of model calibration is the distribution of euro area countries into groups, or clusters, that are homogeneous in terms of their exposure to systematic risk factors. We do this by splitting EFSF shareholders into credit rating groups as follows.

- Cluster 1: comprises the four best credits in the euro area, ie, Germany and Netherlands, both AAA, and Austria and Finland, which are technically rated AA+ but market risk pricing puts them closer to Germany.

---

10To reduce dimensionality, we consider only those countries whose share in the EFSF guarantees is greater than 1%.
Cluster 2: comprises France and Belgium, which are a notch riskier than the best rated peers, owing largely to the sluggish pace of implementing structural reforms, but still very safe, consistent with an AA rating.

Cluster 3: includes Estonia and Slovakia, the two most recent euro area entrants, and Ireland; these countries are still relatively safe, but geopolitical factors (Estonia, Slovakia) or prior recourse to EFSF assistance (Ireland) renders them a notch riskier than peers from Clusters 1 and 2.

Cluster 4: euro area “enfants terribles” Spain, Italy, Portugal and Greece, of which three have already taken recourse to EFSF financing and one (Greece) has recently had its debt restructured (and, in 2015, defaulted on its obligations to the International Monetary Fund); this is the most risky group.

We now associate with each cluster its own systematic risk driver, modeled as a Poisson process, which triggers the default of all cluster members and guarantors in the more junior clusters. Note, however, that this is still a one-factor version of the model as per (4.1), consistent with the fact that there is only one order in which countries are expected to default due to systematic causes. This simple idea lets us strip CDS-implied hazard rates into systematic components (made up by the sum of default intensities of each cluster shock) and idiosyncratic components (defined as residual between CDS-implied intensities and systematic components). The result of this process, using sovereign CDS curves as of February 19, 2016, is shown in Figure 5.

Having defined the default structure in the economy, we can simulate the jumps of the respective Poisson processes and estimate default correlation between euro area sovereigns for a given time horizon. The results – based on 10,000 Monte Carlo runs – are shown in Table 4. Average correlations in Cluster 1 are clearly much higher than in Cluster 4. In particular, in our model Greece has only about 10% default correlation with Portugal and 11% correlation with Italy and Spain, even though the correlation between Portugal and Spain is 40%, and Portugal and Italy is 82%. This pattern reflects a relatively high share (70%) of idiosyncratic default factors in the calibrated hazard rates.

The same Monte Carlo simulation engine can be used to price an EFSF bond as a collection of credit-risky cashflows guaranteed by

(i) the program countries (Greece, Ireland, Portugal) and

(ii) EFSF shareholders (Table 3).

As an example, we consider a €5 billion EFSF bond issued in April 2014 and maturing in June 2021 (ie, with a residual maturity of slightly more than five years as of February
TABLE 4 Simulated default correlations (in percent) for EFSF guarantors and borrowers (two years, 10,000 Monte Carlo runs).

<table>
<thead>
<tr>
<th></th>
<th>GR</th>
<th>PT</th>
<th>IT</th>
<th>ES</th>
<th>EE</th>
<th>IE</th>
<th>SK</th>
<th>BE</th>
<th>FR</th>
<th>AT</th>
<th>FI</th>
<th>NL</th>
<th>DE</th>
</tr>
</thead>
<tbody>
<tr>
<td>GR</td>
<td>100</td>
<td>10</td>
<td>11</td>
<td>11</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PT</td>
<td>10</td>
<td>100</td>
<td>39</td>
<td>41</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>15</td>
<td>15</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>IT</td>
<td>11</td>
<td>39</td>
<td>100</td>
<td>82</td>
<td>42</td>
<td>40</td>
<td>43</td>
<td>31</td>
<td>34</td>
<td>19</td>
<td>19</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>ES</td>
<td>11</td>
<td>41</td>
<td>82</td>
<td>100</td>
<td>53</td>
<td>49</td>
<td>54</td>
<td>39</td>
<td>43</td>
<td>24</td>
<td>24</td>
<td>25</td>
<td>26</td>
</tr>
<tr>
<td>EE</td>
<td>4</td>
<td>20</td>
<td>53</td>
<td>100</td>
<td>83</td>
<td>99</td>
<td>65</td>
<td>80</td>
<td>41</td>
<td>44</td>
<td>46</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>IE</td>
<td>4</td>
<td>19</td>
<td>40</td>
<td>49</td>
<td>83</td>
<td>100</td>
<td>85</td>
<td>58</td>
<td>68</td>
<td>36</td>
<td>38</td>
<td>39</td>
<td>41</td>
</tr>
<tr>
<td>SK</td>
<td>4</td>
<td>20</td>
<td>43</td>
<td>54</td>
<td>99</td>
<td>85</td>
<td>100</td>
<td>66</td>
<td>81</td>
<td>42</td>
<td>44</td>
<td>47</td>
<td>49</td>
</tr>
<tr>
<td>BE</td>
<td>3</td>
<td>15</td>
<td>31</td>
<td>39</td>
<td>65</td>
<td>58</td>
<td>66</td>
<td>100</td>
<td>79</td>
<td>42</td>
<td>44</td>
<td>46</td>
<td>48</td>
</tr>
<tr>
<td>FR</td>
<td>0</td>
<td>15</td>
<td>34</td>
<td>43</td>
<td>80</td>
<td>68</td>
<td>81</td>
<td>79</td>
<td>100</td>
<td>50</td>
<td>54</td>
<td>58</td>
<td>61</td>
</tr>
<tr>
<td>AT</td>
<td>2</td>
<td>9</td>
<td>19</td>
<td>24</td>
<td>41</td>
<td>36</td>
<td>42</td>
<td>42</td>
<td>50</td>
<td>100</td>
<td>64</td>
<td>71</td>
<td>74</td>
</tr>
<tr>
<td>FI</td>
<td>1</td>
<td>9</td>
<td>19</td>
<td>24</td>
<td>44</td>
<td>38</td>
<td>44</td>
<td>44</td>
<td>54</td>
<td>64</td>
<td>100</td>
<td>78</td>
<td>83</td>
</tr>
<tr>
<td>NL</td>
<td>1</td>
<td>9</td>
<td>20</td>
<td>25</td>
<td>46</td>
<td>39</td>
<td>47</td>
<td>46</td>
<td>58</td>
<td>71</td>
<td>78</td>
<td>100</td>
<td>94</td>
</tr>
<tr>
<td>DE</td>
<td>1</td>
<td>9</td>
<td>20</td>
<td>26</td>
<td>48</td>
<td>41</td>
<td>49</td>
<td>48</td>
<td>61</td>
<td>74</td>
<td>83</td>
<td>94</td>
<td>100</td>
</tr>
</tbody>
</table>

FIGURE 6 EFSF bond price as a function of different recovery rate assumptions.

2016) with a fixed annual coupon of 1.375%. Pricing is relatively straightforward and consists in simulating jumps of the Poisson processes associated with each factor. This generates the default times of the respective euro area member states, which in turn can be used to determine the coupon payments on the bond in each scenario. If no defaults occur by coupon payment date $T$, the coupon is paid in full. If any of the program countries (Ireland, Portugal, Greece) defaults, then the guarantors make up for the shortfall, either in full or just to the tune of the chosen recovery rate. Using,
as before, 10,000 Monte Carlo runs and a recovery rate of 0.40 produces a price of 108.788 versus 107.705 quoted by the market as of February 19, 2016. The simulation can be easily adjusted to stress test the calculations for different values of the recovery rate, as shown in Figure 6.

4.4 Calibration to iTraxx Europe

As a final example we show how our model can be applied to pricing CDOs. Our focus will be on synthetic CDO tranches, the market for which has steadily recuperated after the 2008–9 financial crisis, although the approach is general enough to cater also for typical funded structures such as CLOs or ABSs. Despite a marked change in pricing relative to the pre-crisis period (ie, a considerably higher implied correlations and higher spread compensation for mezzanine and senior investors), pricing methodology itself has not changed materially and industry practice still relies on Gaussian copula which is the default valuation model, eg, in the Bloomberg CDO pricer. However, as we have argued above, the problem with the Gaussian copula is that it lacks any real dynamics and does not allow the direct modeling of the concentration of losses in time. This is an important shortcoming, since the concentration of losses itself is not enough to explain the spectacular collapse of the CDO market in 2008–9. Indeed, data compiled by JP Morgan indicates that the average default rate in the universe of credits used in synthetic CDO portfolios has been a mere 2.5% since December 2006 (with an average downgrade of 1.5 notches in each name). In contrast, the large mark-to-market losses that led many investors to exit the market and ultimately led to a cessation of issues in 2008–9 were brought about by a concentration of losses over time, albeit across a variety of portfolios. We have seen above that our model fares significantly better in capturing this “temporal” dimension of default correlation. Below we illustrate its capabilities in fitting the actual market CDO tranche prices.

To fix ideas, consider a CDO on a CDS portfolio referencing \( d \) names (in the case of iTraxx Europe \( d = 125 \)) with a uniform deterministic hazard rate \( R \) and loss-given-default \( \text{LGD} = 1 - R \). To facilitate presentation, assume also that the notional of each obligor is \( \text{LGD}^{-1} \). For any tranche \([A, B]\) with attachment point \( A \) and detachment point \( B \), the protection seller in a CDO pays tranched loss increments at defaults. In exchange, the protection buyer pays an upfront payment \( U \) along with a periodic premium (spread) on dates \( T_1, T_2, \ldots, T_N \). The survival amount associated with the

---

11 In practice some differences in the pool of underlying assets (CDS in synthetic CDOs versus loans in CLOs) are likely to affect the relative pricing of structured products. For example, CLOs are affected by prepayment risk, whereas there is no embedded optionality in synthetic CDOs. Another important consideration is that CLOs are managed and tend to have higher recovery rates.
given tranche $S_{AB}$ is given by

$$S_{AB}(t) = \left(\frac{Bd}{LGD} - N(t)\right)^+ - \left(\frac{Ad}{LGD} - N(t)\right)^+,$$

where, in line with notation introduced in Section 3.1, $N(t)$ counts the number of defaults, ie, $N(t) = \text{card}(\tau_i < t)$ for $i = 1, 2, \ldots, d$. This immediately leads to the expected survival amount

$$E(S_{AB}(t)) = E\left(\frac{Bd}{LGD} - N(t)\right)^+ - E\left(\frac{Ad}{LGD} - N(t)\right)^+$$

$$= \sum_{j=0}^{n(B)} \mathbb{P}(N(t) = j)\left(\frac{Bd}{LGD} - j\right) - \sum_{j=0}^{n(A)} \mathbb{P}(N(t) = j)\left(\frac{Ad}{LGD} - j\right),$$

where

$$n(k) = \max\left\{n \in \mathbb{N} : n < \frac{kd}{LGD}\right\}.$$

The breakeven spread $K_{AB}$ on a tranche with no upfront payment is found by solving the following equation:

$$K_{AB} \sum_{i=1}^{N} \delta_i D(0, T_i)E(S_{AB}(T_i)) + \int_0^{T_N} D(0, t)E(S_{AB}(dt)) = 0,$$ (4.10)

where $\delta_i = T_i - T_{i-1}$ and $D(s, t)$ is the discount factor evaluated at $s$ for time $t$.

It is clear that the price of CDO tranches is determined essentially by the default-counting process $N(t)$. In (3.6)–(3.8) we have provided a very rough closed-form approximation that relies crucially on the breakdown of hazard rates into systematic and idiosyncratic components, with the systematic factor $Z$ represented by an ordered family $Z_1 \leq Z_2 \leq \cdots \leq Z_d$ (see Definition 3.3). A more sophisticated approximation – one that we resort to in the calibration below – consists in assuming that the process counting all systematic defaults is Markovian, so that, with $d \geq n > i$,

$$\mathbb{P}(N^{\text{sys}}(T_{j+1}) = n | N^{\text{sys}}(T_j) = i)$$

$$= \mathbb{P}(N^{\text{sys}}(T_{j+1}) \geq n) - \mathbb{P}(N^{\text{sys}}(T_{j+1}) \geq n + 1)$$

$$= \mathbb{P}(Z_n < T_{j+1} < Z_{n+1})$$

$$= \exp\left(\int_0^{T_{j+1}} \lambda^{\text{sys}}_{n+1}(s) \, ds\right) - \exp\left(\int_0^{T_{j+1}} \lambda^{\text{sys}}_{n}(s) \, ds\right).$$
where $Z_t$ are constructed as in (4.1). If we furthermore assume that all idiosyncratic hazard rates are equal,\textsuperscript{12} ie, that

$$\mathbb{P}(Y_i \leq t) = \mathbb{P}(Y_1 \leq t) = 1 - \exp \left( \int_0^t \lambda^{\text{idio}}_i(s) \, ds \right) \quad \text{for } i = 1, \ldots, d,$$

then the process counting all idiosyncratic defaults $N_0(t)$ is also Markovian with transition probability given by

$$\mathbb{P}(N_0(T_j+1) = n \mid N_0(T_j) = i) \approx \begin{cases} 
(d - i)(\mathbb{P}(Y_1 \leq T_j+1) - \mathbb{P}(Y_1 \leq T_j)) & \text{for } n = i + 1, \\
0 & \text{for } n < i \text{ and } n > i + 1, \\
1 - (d - i)(\mathbb{P}(Y_1 \leq T_j+1) - \mathbb{P}(Y_1 \leq T_j)) & \text{for } n = i.
\end{cases}$$

We now apply the methodology presented above to calibrate the model to iTraxx Europe series 24 as of February 5, 2016 (Table 5). The general idea behind calibration mimics that for pricing EFSF. However, to make use of Markovian approximations we assume uniform idiosyncratic hazard rates. To reduce dimensionality we split the 125 names making up the index into clusters, and assign to each cluster its own systematic risk driver, the arrival of which triggers the default of all names in that cluster as well as those in more junior clusters. We found that, to fit the entire CDO structure simultaneously for three maturities (3Y, 5Y, 7Y), five clusters were needed, comprising 6, 9, 12, 21 and 125 obligors, respectively. The recovery rate was taken to be a standard 40%. The calibration errors in upfront payment (for tranches 0–3% and 3–6%) and quoted spreads (for tranches 6–12% and 12–100%) are shown in Figure 7, which reveals a very good fit, well within bid–ask spreads.

An additional virtue of our pricing approach is that, once the disaggregation of hazard rates into idiosyncratic and systematic components has been performed, it

\textsuperscript{12}This assumption is validated by our finding that the degree of heterogeneity among idiosyncratic hazard rates of individual names does not materially affect CDO prices. What matters instead is the sum of the idiosyncratic hazard rates.

<table>
<thead>
<tr>
<th>Tranche</th>
<th>3Y</th>
<th>5Y</th>
<th>7Y</th>
<th>Average bid–ask spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–3%</td>
<td>35.35%</td>
<td>53.50%</td>
<td>64.25%</td>
<td>2 percentage points</td>
</tr>
<tr>
<td>3–6%</td>
<td>5.13%</td>
<td>14.75%</td>
<td>23.75%</td>
<td>2 percentage points</td>
</tr>
<tr>
<td>6–12%</td>
<td>107.49%</td>
<td>174.59%</td>
<td>223.78%</td>
<td>20bps</td>
</tr>
<tr>
<td>12–100%</td>
<td>28.00%</td>
<td>53.25%</td>
<td>73.00%</td>
<td>10bps</td>
</tr>
</tbody>
</table>

\[\text{TABLE 5} \quad \text{iTraxx Europe series 24 pricing as of February 5, 2016.}\]
becomes possible to gauge the extent to which expected losses (EL) in each tranche are driven by systematic and idiosyncratic factors. To see this, note that by (4.10) we can express the credit spread for the entire index (tranche 0–100%) as

$$K_{0-100} \sum_{i=1}^{N} \delta_i D(0, T_i)(125 - \mathbb{E}(N(T_i))) = 0.6\text{EL},$$

where

$$\text{EL} = \int_0^{T_N} D(0, t)\mathbb{E}(N(dt)) = \sum_{i=1}^{125} \int_0^{T_N} D(0, t)\lambda_i(t)(1 - F_i(t)) \, dt$$
with \( F_i(t) = \mathbb{P}(\tau_i < t) \) being the distribution function describing the default times \( \tau_1, \ldots, \tau_{125} \) of index names. In the model, \( \lambda_i(t) = \lambda_i^{\text{idio}}(t) + \lambda_i^{\text{sys}}(t) \), which leads to the following representation of expected losses disaggregated into idiosyncratic and systematic parts:

\[
EL = EL^{\text{sys}} + EL^{\text{idio}} \\
= \sum_{i=1}^{125} \int_0^{T_N} D(0, t)\lambda_i^{\text{sys}}(t)(1 - F_i(t)) \, dt \\
+ \sum_{i=1}^{125} \int_0^{T_N} D(0, t)\lambda_i^{\text{idio}}(t)(1 - F_i(t)) \, dt.
\]

Hence, assuming that index spreads reflect approximately expected losses, we can use model-based decomposition of hazard rates to gain insight as to the relative weight of idiosyncratic and systematic loss drivers. The results of such decomposition for the iTraxx Europe index are presented in Table 6, which confirms our basic intuition: the longer the maturity, the lower the contribution of idiosyncratic factors to expected losses and the greater the contribution of the fatal shock scenario.

### 5 CONCLUSIONS

Our goal in the foregoing analysis was to present an intuitive, analytically tractable and flexible model of default correlation that could serve as a viable alternative to copula models. In the model, firms’ defaults are driven by both firm-specific and economy-wide shocks. However, in contrast to previous approaches, the systematic factor is represented as a sequence of increasing random variables. This characterization can be interpreted as an invisible chain of dependencies running through the whole economy, whereby a systematic default of \( i \)th obligor causes defaults of all more systematically risky names. The model is arbitrage-free and allows us to construct...
an almost arbitrarily rich correlation structure. A number of examples – from simple repos to CDS portfolios, EFSF bonds and CDOs – was provided to show how the model can be applied and calibrated in different contexts.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES


Research Paper

Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default

Han Sheng Sun¹ and Zi Jin²

¹Global Risk Management, Scotiabank, 14th Floor, 4 King Street West, Toronto, ON M5H 1B6, Canada; email: hansheng.sun@scotiabank.com
²Corporate Model Risk, Wells Fargo Bank, 6th Floor, 550 California Street, San Francisco, CA 94104, USA; email: zi.jin@wellsfargo.com

(Received December 23, 2014; revised October 29, 2015; accepted February 3, 2016)

ABSTRACT

In credit risk modeling, banks and insurance companies routinely use a single model for estimating key risk parameters. Combining several models to make a final prediction is not often considered. Using an ensemble or a collection of models rather than a single model can improve the accuracy and robustness of prediction results. In this study, we investigate two well-established ensemble learning methods (stochastic gradient boosting and random forest) and propose two new ensembles (ensemble by partial least squares and bag-boosting) in the application of predicting the loss given default. We demonstrate that an ensemble approach significantly increases the discriminatory power of the model compared with a single decision tree. In addition, the ensemble learning methods can be applied directly to predicting the exposure at default and probability of default with some simple modifications. The proposed approaches introduce a novel modeling framework that banks and other financial institutions can use to estimate and validate credit risk parameters based on the internal data of different portfolios. Moreover, the proposed approaches can be readily extended to general portfolio risk modeling in the areas of regulatory capital and

Corresponding author: H. S. Sun

Print ISSN 1744-6619 | Online ISSN 1755-9723
Copyright © 2016 Incisive Risk Information (IP) Limited

To subscribe to a Risk Journal visit Risk.net/subscribe or email corpsubs@risk.net
1 INTRODUCTION

In credit risk modeling, typically a single model, either a regression model or a decision tree, is used to estimate three key risk parameters: probability of default (PD), loss given default (LGD) and exposure at default (EAD). The approach to combining several (tree) models in making final predictions is not often considered. Using an ensemble or a collection of models rather than a single model can improve the accuracy and robustness of prediction results. Ensemble learning methods have many applications in the field of data mining and artificial intelligence. However, their applications to credit risk modeling have rarely been explored in the literature, and empirical results are scarce.

The general idea of ensemble modeling is to combine several individual models, base learners, in a reasonable way to obtain a final model that outperforms each of them individually.

Basically, an ensemble model consists of three components:

(1) training data, consisting of a list of explanatory variables representing the input features and corresponding target variables;

(2) a few base learners, ie, the models used for basic predicting; and

(3) an ensemble algorithm, ie, a procedure to combine base learners into a strong ensemble.

In this study, we investigate ensemble learning methods on credit risk modeling in detail, focusing specifically on predicting LGD. We apply two well-established ensemble methods – stochastic gradient boosting (Friedman 2001) and random forest (Breiman 2001) – to LGD prediction. Moreover, to further improve the model performance, we propose two novel ensemble methods: partial least squares (PLS) and bag-boosting. Our study is conducted on two empirical data sets: retail exposure based on the data of a large North American bank; and commercial and capital market exposure using data from Moody’s Ultimate Recovery Database (URD). All four ensemble methods we investigate significantly outperform the single regression tree for both retail and wholesale exposures; moreover, these methods can improve their performance further. The rationale for the proposed ensemble methods is discussed
in detail and can help readers to develop ensemble methods tailored to their specific modeling needs.

Compared with conventional approaches, the ensemble methods are capable of making more accurate, robust and stable predictions, with appreciable simplicity. These desired properties allow applications in many credit risk modeling areas. For instance, the methods presented in this paper can be readily applied to EAD and PD predictions with some simple modifications. Together with LGD, the accuracy improvements in key risk parameters can make fundamental contributions to financial institutions in the areas of regulatory capital and economic capital management, loss forecasting, stress testing and pre-provision net revenue (PPNR) projections. Despite the variety of applications, we focus here on the application of ensemble methods to predicting LGD.

The LGD parameter measures the percentage loss incurred over the total exposure if an obligor defaults, typically defined by

\[
\text{LGD} = 1 - \text{recovery},
\]

where

\[
\text{recovery} = \frac{\text{recovered amount}}{\text{total exposure at default}}.
\]

The LGD data could be bimodal, containing observations from near complete recovery and low recovery, or skewed and heavy tailed, depending on the nature of the recovery. Accurately predicting the LGD is a challenging problem. There is substantially less published work on LGD modeling than there is on PD modeling. Depending on the availability of modeling samples, both parametric and nonparametric models can be considered.

Beta regression and lognormal regression are common parametric choices. For example, Gupton and Stein (2005) used Beta transformation, and Dermine and Neto de Carvalho (2006) and Caselli et al (2008) considered a log transformation before fitting a linear regression model. However, it is often challenging to make proper distributional assumptions, because the LGD of the underlying population usually shifts with economic conditions, and predictive variables (features) are often missing data, which makes it hard to develop an accurate and robust parametric model. Nevertheless, if properly constructed, parametric models can make strong predictions. On the other hand, a nonparametric model makes no assumption on the probability distribution of LGD, is less sensitive to data issues such as missing values and generally provides robust predictions. Nonparametric models usually require a larger data set to make predictions at the same confidence level as parametric models. This is often not an issue for retail LGD modeling, where data sets are sufficiently large, but may be challenging for commercial exposures with small sample sizes.
Bastos (2010) recommended using regression trees for LGD modeling, and Loterman et al (2012) compared twenty-four techniques, including ordinary least squares regression, Beta regression, robust regression, ridge regression, regression splines, neural networks (NNs) and support vector machines (SVMs), on six loss data sets, and concluded that nonparametric techniques such as NNs and kernel SVMs generally perform better. More recently, Bastos (2014) recommended using an ensemble method of bootstrap aggregation (“bagging”) to predict LGD and showed that it can improve prediction accuracy by over 40% compared with the linear regression model, and by over 20% compared with the regression tree.

Our findings are generally consistent with Bastos’s results. However, our paper uses two state-of-the-art ensemble methods: stochastic gradient boosting and random forest, and proposes two new ensemble algorithms that are not available in the standard machine learning literature. Furthermore, we provide a detailed discussion and a performance comparison on both retail and wholesale exposures.

The paper proceeds as follows. Section 2 provides detailed descriptions of the data sets we use. Section 3 establishes the baseline model, discusses two popular ensemble methods and proposes two new ensemble algorithms that combine the elements of the existing ensemble approaches. The details of the algorithms and implementations for each ensemble method are provided in the online appendixes. Section 4 presents model performance comparisons of the five methods (baseline and four ensemble methods) on the two data sets. Section 5 discusses model interpretability of the ensemble methods by focusing on the economic ideas behind the input features. Section 6 summarizes our findings and examines natural extensions.

2 DATA DESCRIPTIONS

Two data sets are used to illustrate the advantage of ensemble modeling over decision-tree models. The first data set is based on the consumer default history of a Retail Revolver portfolio from a large North American bank (see Section 2.1). The second is taken from Moody’s URD (see Section 2.2).

2.1 Retail Revolver data set

A sample of retail portfolio data containing 3500 observations with 499 predictive features is divided randomly into 60% training samples (2100 accounts) and 40% validation samples (1400 accounts). The input features are behavior and transaction related, and, unsurprisingly, many of the predictors have missing values. The time horizon of default years ranges from October 2005 to October 2011, covering the benign economic conditions from 2005 up to the first half of 2007, the severe stress period of 2007–9 and the moderate stress period of 2009–11. The LGD is defined as...
the workout loss of defaulted loans. The workout period is the twenty-four months after default. The recovery cashflows are discounted to the time of default, with a fixed annual discount rate of 10%, assuming that it reflects the cost of equity of the bank. If a defaulted customer is able to successfully pay off their debt during the subsequent twenty-four-month period, the account is considered “cured” and assigned an LGD of zero. More defaults and higher levels of LGD were observed during the economic downturn. The empirical distribution of the LGD data is shown in Figure 1.

The 499 predictive features reflect available information collected on a loan, which can be categorized into three groups. First, the borrower or loan’s generic information, including borrower information collected at the time of their application, eg, occupation, income, credit history, as well as collateral information. Second, account balance, credit limit, spending and delinquency information collected from the bank’s internal payment system and credit bureau during the twenty-four months prior to default. Third, variables derived to capture accounts’ trend, change, summary, extreme values and other potentially useful features. Although many of these features are highly correlated and have missing values, decision-tree models are robust enough to handle such data issues. To be clear, the comparison of the performance of a single regression tree with that of ensemble methods conducted later in the paper is based on the same comprehensive set of candidate features.
TABLE 1  Moody’s URD data LGD summary. [Table continues on next page.]

(a) By year

<table>
<thead>
<tr>
<th>Year</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
<th>Year</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985</td>
<td>2</td>
<td>0.865</td>
<td>0.001</td>
<td>1998</td>
<td>74</td>
<td>0.607</td>
<td>0.329</td>
</tr>
<tr>
<td>1987</td>
<td>40</td>
<td>0.343</td>
<td>0.310</td>
<td>1999</td>
<td>193</td>
<td>0.457</td>
<td>0.377</td>
</tr>
<tr>
<td>1988</td>
<td>69</td>
<td>0.510</td>
<td>0.390</td>
<td>2000</td>
<td>283</td>
<td>0.530</td>
<td>0.373</td>
</tr>
<tr>
<td>1989</td>
<td>82</td>
<td>0.565</td>
<td>0.346</td>
<td>2001</td>
<td>603</td>
<td>0.554</td>
<td>0.368</td>
</tr>
<tr>
<td>1990</td>
<td>196</td>
<td>0.536</td>
<td>0.373</td>
<td>2002</td>
<td>764</td>
<td>0.546</td>
<td>0.342</td>
</tr>
<tr>
<td>1991</td>
<td>277</td>
<td>0.435</td>
<td>0.343</td>
<td>2003</td>
<td>369</td>
<td>0.361</td>
<td>0.333</td>
</tr>
<tr>
<td>1992</td>
<td>129</td>
<td>0.372</td>
<td>0.388</td>
<td>2004</td>
<td>200</td>
<td>0.282</td>
<td>0.312</td>
</tr>
<tr>
<td>1993</td>
<td>101</td>
<td>0.449</td>
<td>0.388</td>
<td>2005</td>
<td>194</td>
<td>0.306</td>
<td>0.271</td>
</tr>
<tr>
<td>1994</td>
<td>66</td>
<td>0.343</td>
<td>0.348</td>
<td>2006</td>
<td>70</td>
<td>0.299</td>
<td>0.344</td>
</tr>
<tr>
<td>1995</td>
<td>118</td>
<td>0.425</td>
<td>0.340</td>
<td>2007</td>
<td>37</td>
<td>0.212</td>
<td>0.281</td>
</tr>
<tr>
<td>1996</td>
<td>66</td>
<td>0.410</td>
<td>0.366</td>
<td>2008</td>
<td>119</td>
<td>0.303</td>
<td>0.363</td>
</tr>
<tr>
<td>1997</td>
<td>60</td>
<td>0.394</td>
<td>0.370</td>
<td>2009</td>
<td>37</td>
<td>0.483</td>
<td>0.436</td>
</tr>
</tbody>
</table>

(b) By instrument type

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
<th>Instrument</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Junior subordinated Bonds</td>
<td>66</td>
<td>0.801</td>
<td>0.299</td>
<td>Revolver</td>
<td>845</td>
<td>0.223</td>
<td>0.268</td>
</tr>
<tr>
<td>Subordinated Bonds</td>
<td>365</td>
<td>0.732</td>
<td>0.326</td>
<td>Term loan</td>
<td>754</td>
<td>0.299</td>
<td>0.315</td>
</tr>
<tr>
<td>Senior subordinated Bonds</td>
<td>439</td>
<td>0.714</td>
<td>0.315</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior unsecured Bonds</td>
<td>1121</td>
<td>0.558</td>
<td>0.339</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Senior secured Bonds</td>
<td>559</td>
<td>0.415</td>
<td>0.318</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.2 Moody’s Ultimate Recovery Database

A commercial/capital market data sample is obtained from Moody’s URD. The data set contains 4152 defaulted instruments, including bonds and loans, covering the period from 1985 to 2009, with a total of 324 obligors of US nonfinancial holding corporations. Table 1 presents an LGD summary of the data. The data set is divided randomly into two subsets: a 70% training population and 30% for out-of-sample validation by issuers. Since the number of the issuers’ defaulted instruments (loans and bonds) may vary, the final training data set contains 2813 instruments, whereas the validation data set contains 1136 instruments.

Similarly to the retail data set, the recovery cashflows are discounted to the time of default with a fixed annual discount rate of 10%. The distribution of the overall
TABLE 1  Continued.

(c) By Moody's industry group

<table>
<thead>
<tr>
<th>Industry</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
<th>Industry</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Automotive</td>
<td>141</td>
<td>0.499</td>
<td>0.372</td>
<td>Leisure &amp; entertainment</td>
<td>228</td>
<td>0.395</td>
<td>0.381</td>
</tr>
<tr>
<td>Chemicals</td>
<td>55</td>
<td>0.432</td>
<td>0.357</td>
<td>Manufacturing</td>
<td>362</td>
<td>0.421</td>
<td>0.333</td>
</tr>
<tr>
<td>Construction</td>
<td>52</td>
<td>0.487</td>
<td>0.363</td>
<td>Media</td>
<td>258</td>
<td>0.364</td>
<td>0.324</td>
</tr>
<tr>
<td>Consumer products</td>
<td>342</td>
<td>0.383</td>
<td>0.344</td>
<td>Metals &amp; mining</td>
<td>128</td>
<td>0.483</td>
<td>0.359</td>
</tr>
<tr>
<td>Distribution</td>
<td>519</td>
<td>0.522</td>
<td>0.375</td>
<td>Natural products</td>
<td>85</td>
<td>0.206</td>
<td>0.211</td>
</tr>
<tr>
<td>Energy</td>
<td>467</td>
<td>0.340</td>
<td>0.313</td>
<td>Packaging</td>
<td>23</td>
<td>0.250</td>
<td>0.329</td>
</tr>
<tr>
<td>Environment</td>
<td>44</td>
<td>0.763</td>
<td>0.258</td>
<td>Services</td>
<td>322</td>
<td>0.454</td>
<td>0.381</td>
</tr>
<tr>
<td>Health care</td>
<td>148</td>
<td>0.484</td>
<td>0.380</td>
<td>Technology</td>
<td>134</td>
<td>0.464</td>
<td>0.362</td>
</tr>
<tr>
<td>Holding company</td>
<td>14</td>
<td>0.977</td>
<td>0.045</td>
<td>Telecommunications</td>
<td>464</td>
<td>0.604</td>
<td>0.386</td>
</tr>
<tr>
<td>Industrials</td>
<td>46</td>
<td>0.461</td>
<td>0.357</td>
<td>Transportation</td>
<td>303</td>
<td>0.554</td>
<td>0.320</td>
</tr>
</tbody>
</table>

(c) By collateral type

<table>
<thead>
<tr>
<th>Collateral</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
<th>Collateral</th>
<th>Obs</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounts receivable</td>
<td>41</td>
<td>0.164</td>
<td>0.233</td>
<td>Inventory and accounts receivable</td>
<td>116</td>
<td>0.079</td>
<td>0.104</td>
</tr>
<tr>
<td>All assets</td>
<td>1156</td>
<td>0.241</td>
<td>0.280</td>
<td>Most assets</td>
<td>22</td>
<td>0.223</td>
<td>0.199</td>
</tr>
<tr>
<td>All noncurrent assets</td>
<td>30</td>
<td>0.359</td>
<td>0.265</td>
<td>Oil and gas properties</td>
<td>8</td>
<td>0.082</td>
<td>0.082</td>
</tr>
<tr>
<td>Capital stock</td>
<td>179</td>
<td>0.378</td>
<td>0.294</td>
<td>Other</td>
<td>2</td>
<td>0.443</td>
<td>0.523</td>
</tr>
<tr>
<td>Cash</td>
<td>14</td>
<td>0.103</td>
<td>0.147</td>
<td>PP&amp;E</td>
<td>164</td>
<td>0.402</td>
<td>0.299</td>
</tr>
<tr>
<td>Equipment</td>
<td>120</td>
<td>0.600</td>
<td>0.296</td>
<td>Real estate</td>
<td>25</td>
<td>0.286</td>
<td>0.329</td>
</tr>
<tr>
<td>Guarantees</td>
<td>16</td>
<td>0.077</td>
<td>0.089</td>
<td>Second lien</td>
<td>140</td>
<td>0.420</td>
<td>0.348</td>
</tr>
<tr>
<td>Intellectual property</td>
<td>5</td>
<td>0.569</td>
<td>0.313</td>
<td>Third lien</td>
<td>30</td>
<td>0.627</td>
<td>0.362</td>
</tr>
<tr>
<td>Intercompany debt</td>
<td>1</td>
<td>0.878</td>
<td>0.315</td>
<td>Unsecured</td>
<td>2060</td>
<td>0.628</td>
<td>0.341</td>
</tr>
<tr>
<td>Inventory</td>
<td>18</td>
<td>0.169</td>
<td>0.152</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SD denotes standard deviation. PP&E denotes property, plant and equipment.

LGD exhibits a typical bimodal pattern. However, the LGDs of loan instruments are clearly lower than those of the bond instruments, which reflects the seniority of loan instruments, as presented in Figures 2 and 3.

Unlike the retail data sets, only eighteen input features are available for modeling, including debt seniority, interest rate information and industrial- and collateral-type information. There is no macroeconomic variable, but an economic downturn
indicator is included in the data. Including macroeconomic variables may potentially enrich input features and improve the accuracy of the prediction.

3 CART AND ENSEMBLE METHODS

3.1 Classification and regression tree

The most popular nonparametric method for LGD modeling is the classification and regression tree (CART), first introduced by Breiman et al (1984). The method has various desirable properties: making robust predictions; scaling well with a large data set, requiring little data preparation; and an ability to handle both numerical and categorical variables. Since LGD is a real-valued continuous target, a regression tree is often used.

The regression tree is a widely used effective and robust methodology in predictive modeling. The tree grows by greedy searching for the optimal split points at each node until a stopping criterion – usually a constraint on the depth of the tree and the minimal size of a node – is reached. The optimal split point can be determined by maximizing the reduction of an impurity measure of the target value given the split, or testing the significant difference between the target values of the different branches defined by a candidate split. A popular choice of impurity measure is the average square error; this method of determining the splitting rule is commonly known as variance reduction. In this paper, we use hypothesis testing to determine the optimal
split. The details of the mathematical construct are provided in Appendix 1, available online.

### 3.2 Ensemble learning methods: an overview

An early mathematical foundation of ensemble learning is the jury theorem published by French mathematician Nicolas de Caritat, Marquis de Condorcet in 1875.
The theorem refers to a jury of voters who need to reach a decision by a majority vote. The decision is binary: one of the two outcomes is correct. If each voter has an independent probability $p$ of being correct and the probability of the majority of voters being correct is $E$, then

- $p > 0.5$ implies $E > p$,
- $E$ approaches 1, for all $p > 0.5$, as the number of voters approaches infinity.

The ensemble idea was investigated in the field of predictive modeling as early as 1970s. Tukey (1977) suggested combining two linear regression models, the first for fitting the original data, the second for fitting the residuals. The breakthrough of the ensemble came in 1996, when two important methods were published: adaptive boosting (AdaBoost) and bootstrap aggregating (bagging). Freund and Schapire (1996) introduced the AdaBoost algorithm by showing that a strong classifier in the probably approximately correct (PAC) sense can be generated by combining weak classifiers. Breiman (1996) proposed another ensemble algorithm (bagging), where the weak classifiers are combined in a different way. A comprehensive review of these methods is given by Rokach (2010). The emergence of ensemble learning methods inspired a wave of research in statistical machine learning; its superior performance in both predictive accuracy and robustness has been witnessed.

The main idea behind the ensemble learning methods is particularly far-reaching: we can make accurate predictions by building an ensemble of models without much fine-tuning of each member of the ensemble, and achieve better performance than carefully fine-tuning a single model. The fundamental reason why an ensemble works is that combining a collection of weak learners in a reasonable way will produce a strong prediction result.

In Section 3.3 and 3.4, we will introduce two well-established state-of-the-art ensemble algorithms: stochastic gradient boosting, a sequential optimization process; and random forest, a parallel development process, and their applications to predicting LGD. In Sections 3.5 and 3.6 we propose two new algorithms leveraging the principles discussed in Section 3.4.

### 3.3 Stochastic gradient boosting

Gradient boosting was first introduced by Friedman (2001) as a general method for function approximation. Gradient boosting is an iterative method of fitting a collection of weak predictors (base learner) to minimize a loss function; all the predictors additively make the final prediction. At each iteration $m$, a new predictor $f_m$ is multiplied by the learning rate parameter $\gamma_m$ and added to the current ensemble predictor $F_{m-1}$. The new predictor $f_m$ is chosen approximately in order to minimize the negative derivative of the cost function, the pseudo-residual. The steps are iterated $M$
times, eventually giving the final ensemble predictor, $F_M$. The detailed algorithm is presented in Appendix 2, available online.

Note that, at the beginning of each iteration, a bootstrapped subset of the whole training data is drawn, and as a result each individual predictor is constructed on a sample of the training data instead of the whole sample; this step was introduced by Friedman (2001) as an important tweak of the original gradient boosting, and he named it stochastic gradient boosting. This modification is inspired by Breiman’s (1996) bagging algorithm and improves the performance.

In terms of base learners, from our experience, in the application of LGD and EAD modeling, the most successful choice is a weak regression tree with a limited depth of between 2 and 4, depending on the size of the data set. The depth may be increased to accommodate a large data set.

### 3.4 Random forest and Breiman’s recipe

The random forest is a method of constructing a collection of classification or regression trees to create the final ensemble predictor. It is proposed as an improvement to the bagging algorithm (Breiman 1996). The random forest algorithm independently trains $B$ trees, each on a bootstrapped subset of the training data. During the training of each individual tree, the splitting rule is decided using a subset of predictive features instead of all possible features. The final prediction is the average prediction from all trees in the forest, if regression trees are used, or the majority vote in the case of the classification tree. The algorithm can be used for both LGD/EAD modeling (regression trees for continuous targets) and PD modeling (classification trees for binary targets). The detailed algorithm and implementation are presented in Appendix 3, available online.

Breiman (2001) proved a remarkable theoretical result: the generalization error of a random forest $\varepsilon_{RF}$ satisfies the following inequality:

$$\varepsilon_{RF} \leq \tilde{\rho}\left(\frac{1 - s^2}{s^2}\right),$$  \hspace{1cm} (3.1)

where $\tilde{\rho}$ is the mean correlation between any two members of the forest, and $s$ is the mean strength of a typical member of the forest. The proof and exact definition of $\tilde{\rho}$ and $s$ are technical, and can be found in Breiman’s original paper. However, the actual bound itself is not tight, and often useless in practice. A good summary of the result and its implication is presented by Zhu (2008).

Nevertheless, Breiman’s theory gives us a recipe for good random forests (ensemble). The theory suggests that a good random forest should have small $\tilde{\rho}$ and large $s$. That is, we should reduce the correlation between the individual base learners.
and make each base learner as accurate as possible. It is interesting to note that the gradient boosting algorithm follows a similar principle. At each iteration the pseudo-residuals become the new targets, ie, \( f_{m+1} \) is built using the residual created by the current ensemble predictor \( F_m \). This step actually reduces the correlation between the consecutive members of the ensemble.

Following this principle, we propose two new ensemble algorithms – ensemble by PLS and bag-boosting – and apply them to LGD modeling in Sections 3.5 and 3.6. These proposed methods aim to increase the power of the random forest and gradient boosting by leveraging Breiman’s inequality (3.1). Both methods use regression trees as the base learner.

### 3.5 Ensemble by partial least squares (PLS)

Recall from inequality (3.1) that the power of a random forest can potentially be increased if the average correlation between the individual trees, \( \tilde{\rho} \), is reduced. Moreover, in the current random forest algorithm, each tree \( f_b \) contributes equally to the final ensemble \( F(x) = \sum_{b=1}^{B} f_b / B \), regardless of its strength. Hence, the aim of the proposed ensemble by partial least squares algorithm is twofold: reducing correlations between the trees, and letting stronger trees make a bigger contribution to the final prediction.

The ensemble by PLS follows a similar structure to the bagging algorithm, with two modifications. First, a random subsetting of the predictive features is introduced when creating the training samples of each individual. The purpose of adding this step is to reduce the correlation between trees in the final ensemble predictor. Second, when aggregating the predictions of individual regression trees, instead of taking a simple average, as in bagging or random forest, we use the PLS to aggregate them. This allows the stronger trees in the ensemble to contribute more to the final prediction while controlling the multicollinearity between individual trees.

Using the predicted results of the individual trees in the forest as the new input features to extract the latent factors with the targets, PLS allows us to effectively control the potential over-contribution of correlated trees and leads to better generalizability of the model. To make an analogy, random forest is similar to voting in a general election and ensemble by PLS is more similar to voting by elected representatives.

In addition, the ensemble by PLS can be applied to predicting PD if we simply substitute the base learner of regression trees with classification trees, and then \( f_b \) becomes an observation of the predicted probability of default; we keep the PLS step unchanged. The detailed algorithm and implementation are presented in Appendix 4, available online.
3.6 Bag-boosting

The aim of the proposed bag-boosting algorithm is to improve the performance of an ensemble by increasing the accuracy of each individual predictor; that is, improving the strength $s$ of base learners in inequality (3.1). The algorithm combines the underlying ideas of the random forest and gradient boosting methods. Bag-boosting follows a similar general structure to the bagging algorithm, but instead of aggregating regression trees it aggregates gradient-boosted regression trees. By boosting each regression tree on a bootstrapped sample, we are able to increase the accuracy of each individual predictor without increasing the correlations between them, which essentially leads to smaller generalization error for the aggregated ensemble predictor given by Breiman’s inequality (3.1).

In practice, we only use five boosting steps for each gradient-boosted regression tree, and obtain a general improvement in performance, which will be illustrated in Section 4. The algorithm is essentially an ensemble of ensemble models. The detailed algorithm and its implementation are presented in Appendix 5, available online.

We make no theoretical claim for dominance in terms of the performance of the proposed methods compared with that of the ensemble methods discussed in Sections 3.3 and 3.4, since the optimal performance of different methods can be both data and parameter dependent. However, the newly proposed methods do often have advantages when we need to make robust predictions in credit risk modeling applications, as we will demonstrate in Section 4 with two example data sets.

4 MODEL PERFORMANCE COMPARISON

In this section, we apply the four ensemble models discussed in Section 3 to credit risk LGD modeling using the two data sets described in Section 2. We compare their performance with a baseline single regression tree model, develop the principles of ensemble learning methods and illustrate how to design effective and powerful ensemble algorithms from a practitioner’s standpoint. Sections 4.1–4.4 provide model performance of each ensemble method, and Section 4.5 presents the summary results.

Model performance is compared in terms of discriminatory power, as this is of greater interest to us than the absolute accuracy of the prediction.\(^1\) To measure a model’s discriminatory power, the cumulative accuracy profile curve and its summary statistics’ accuracy ratio are typically used. A detailed explanation of this method can be found in Sobehart et al (2000). For LGD predicting, which is continuous rather than...
than binary, the accuracy ratio of the revised cumulative accuracy profile (RCAP) is used. This measure is largely consistent with the receiver operating characteristic curve and maximum Kolmogorov–Smirnov distance in making ordinal comparisons. In addition, for the results presented in this paper, the RCAP ratio is also consistent with accuracy measures such as the root mean-squared error, since regression trees are used with no variable transformation; that is, the model with the highest RCAP ratio has the smallest root mean-squared error. A hypothesis test (\( F \)-test) is used to determine the optimal split for a regression tree and base learners. The reader may refer to Appendix 1 online for details of the \( F \)-test.

For the baseline model, an optimal single regression tree can achieve the best RCAP ratios of 0.364 and 0.605 on the validation sample for the Retail Revolver data and Moody’s URD data, respectively. See Appendix 1 for algorithm details and parameter settings.

4.1 Stochastic gradient boosting

The base learner of the stochastic gradient boosting is a shallow regression tree. See Appendix 2 online for algorithm details and parameter settings. At each iteration, a bootstrapped data set is generated from the training sample via sampling with replacement. A new predictor is iteratively produced by minimizing the negative derivative of the squared loss function. Figure 4 illustrates the model performance per iteration step. Note that, for the Retail Revolver data set, model performance starts to decrease after fifty iterations. This means the algorithm can occasionally over-fit the training sample in practice, even though it usually makes robust predictions.

The experiment shows (Figure 4) that we can start from a very weak base learner, which has a low/medium discriminatory power (Retail Revolver RCAP = 0.291; Moody’s URD: RCAP = 0.601) on the validation data set, and boost its performance (to 0.419 for the Retail Revolver data set and 0.664 for the Moody’s URD data set) after fifty iterations, which represents a 15% improvement over a single regression tree (RCAP = 0.364) for the Retail Revolver and 10% increase for the Moody’s URD (RCAP = 0.605). Hence, rather than fine-tuning a single tree, we could (just) build a simple decision tree and iterate it a reasonable number of times by applying stochastic gradient boosting and achieve greater discriminatory power.

4.2 Random forest

The random forest algorithm trains many regression trees independently, each tree being on a bootstrapped subset of the training data. The final prediction is made collectively by all trees. See Appendix 3 online for algorithm details and parameter settings. Figure 5 shows model performance in terms of different sizes of the forest. The results indicate that the discriminatory power of the forest increases with the
FIGURE 4 Performance of stochastic gradient boosting: (a) Retail Revolver data and (b) Moody’s URD data.

forest size. More importantly, comparing a forest with 50 trees with one with 150 trees, we see that increasing the size of the model does not lead to any over-fitting; the RCAPs on the validation sample remain stable.
In practice, the individual trees in the forest should be constructed to be as strong as possible (here, we choose a tree of three layers for Retail Revolver, and five for Moody’s URD), in contrast to the gradient boosting algorithm, where the choice of the base learner should be relatively weak. Compared with boosting methods, the random
forest is generally more resistant to over-fitting. Therefore, a random forest should be built with a reasonable size, given acceptable computational cost. We decided to stop at a forest of 150 trees, considering it to have enough performance lift. Moreover, the method is not very sensitive to the choice of parameters, such as the splitting criteria
FIGURE 7 Performance of ensemble by bag-boosting: (a) Retail Revolver data and (b) Moody’s URD data.

(a)

RCAP on validation

Number of boosted trees

0.40

0.35

0.30

0.25

0

10

20

30

40

50

Bag-boosting

Single regression tree

(b)

RCAP on validation

Number of boosted trees

0.68

0.66

0.64

0.62

0.60

0.58

0.56

0

10

20

30

40

50

Bag-boosting

Single regression tree

of individual trees. Depending on the size of input features, modelers can choose the percentage of input features to be used for building the trees in the forest. However, too large a percentage is not encouraged because the generalization error of a random
forest, $\varepsilon_{RF}$, increases with the mean correlation between any two members of the forest. Too many similar input features between trees increases tree correlations. If 100% of input features are used for each tree, the random forest degenerates to the bagging method, which is less powerful.

4.3 Ensemble by PLS

The ensemble by PLS uses partial least squares to aggregate the results from individual trees. See Appendix 4 online for algorithm details and parameter settings. The model performance is presented in Figure 6. Our results show the ensemble by PLS is able to achieve the same performance as random forest by training a smaller number of trees; the RCAP of 0.415 is achieved with 30 trees while random forest requires 150 trees to obtain the same RCAP level. Hence, the ensemble by PLS is computationally efficient, and is recommended if computational cost is of concern. Moreover, the model discriminatory power can be further improved to an RCAP of 0.436, compared with an RCAP equal to 0.415 for the random forest with the same number of trees. This also indicates that increasing the number of trees does not usually lead to over-fitting of the training samples due to the PLS step. However, for the Moody’s URD data, although the method improved the RCAP from 0.605 to 0.663, increasing the number of trees did not lead to a consistent improvement in the discriminatory power. This may be caused by the small number of candidate features available.

4.4 Bag-boosting

Bag-boosting aims to improve the ensemble performance by increasing the accuracy of each base learner. The final predictor is an aggregate of gradient-boosted regression trees. See Appendix 5 online for algorithm details and parameter settings. Figure 7 shows the bag-boosting performance in terms of the number of trees.

The bag-boosting method appears to produce the strongest model of the ensemble methods presented herein. For the Retail Revolver data set, including five boosting steps, this ensemble model achieves a good performance, with an RCAP of 0.420 with only ten members. A better performance, with an RCAP of 0.431, is achieved with fifty members. For the Moody’s URD data set, the ensemble model is able to outperform the single regression tree with only three members and achieves the best performance, with an RCAP equal to 0.676. The method resists over-fitting, as we observe a consistent increase in performance when more boosted trees are included. However, bag-boosting is the most computationally intensive method discussed herein.

4.5 Performance comparisons

In summary, comparing the best performance achieved by all the methods, for the Retail Revolver data set all four ensemble models outperform the single regression
Some notable advantages of using ensemble modeling approaches can be seen.

- Strong model performance.
- Resistance to over-fitting and model robustness: fitting a larger model on the training data rarely leads to over-fitting.
- Foolproof modeling method: performance of the model is not usually sensitive to the choice of parameters (e.g., size of the forest, number of boosting steps), and a wide range of parameters can lead to a strong final model.

In the empirical study, we made a decent effort to ensure fair comparisons, such as using cross-validation for parameter selections, but readers should be aware that our analysis was conducted on two LGD data sets and thus is subject to the constraints of
the nature of the portfolios and powerfulness/structure of the input features as well as data quality.

In practice, the benefit of using an ensemble may vary from case to case. However, as a general methodology, the ensemble learning methods usually give more robust and accurate prediction results. In particular, the richer the data sets, the greater the expectation that ensembles work better. In other words, ensembles provide a good alternative modeling approach when facing challenging and complex credit risk modeling problems. In terms of the choice of ensembles, our experience shows random forest is usually a good start: it is intuitively easy to interpret and straightforward to implement, resists over-fitting and achieves better performance at lower cost. (Stochastic) gradient boosting is worthy of consideration to iteratively improve a tree’s performance instead of simply focusing on fine-tuning it. Moreover, the newly proposed ensemble by PLS and bag-boosting leverage the core principles of ensemble from different perspectives, and are hence able to provide additional improvements to the performance.

5 MODEL INTERPRETABILITY AND PREDICTOR SELECTION

In addition to demonstrating the performance improvement of the ensemble methods, this section discusses the interpretability of the approach and illustrates the economic ideas behind the input features selected by the ensemble methods using the two example data sets.

One common criticism of the ensemble modeling approach is the loss of interpretability: the final ensemble predictor is usually a weighted sum (or some other combination) of trees, which necessarily cannot be represented by a single tree. A solution to mitigate the loss of interpretability is to introduce a measure of the importance of each variable. For a single decision tree \( f_{\text{tree}} \), define the squared importance for predictive variable (feature) \( x_j \) as

\[
\text{Imp}^2_j(f_{\text{tree}}) = \sum_{k=1}^p \hat{d}_k \times I\{\text{split at node } k \text{ is on variable } x_j\},
\]  

(5.1)

where \( p \) is the number of internal nodes (non-leaves), and \( \hat{d}_k \) is the improvement in training misclassification error from making the \( k \)th split. For boosting, we define the squared importance for a variable \( x_j \) by simply averaging the squared importance over all the trees. For random forest, Breiman (2001) suggests using the reduction in out-of-bag error to measure the importance of variables. The out-of-bag samples are the training samples that are not selected by the bootstrapping step of the algorithm. The importance of the variables is also a good measure for selecting variables; we usually set the largest variable’s importance to be 1, and scale all other variables’ importance accordingly.
For simplicity, we base the following discussion on the bagging algorithm. The four ensemble algorithms presented in Section 3 are more complicated, but follow the same principle. Although the interpretability of a model is reduced due to the fact that we lose the ability to readily observe the hierarchical structure and cutoff points for each feature of a regression tree, it is still possible to gain valuable business intuition from an ensemble model to assist key business decision making.

5.1 Retail Revolver data set

Recall from Section 2 that there are a total of 499 input features used to develop the LGD model for the Retail Revolver data set, reflecting the comprehensive information that a bank collects on its customers.

A single regression tree model with the best in-sample validation performance selects seven predictors out of the 499 candidates. Table 4 lists the variables’ descriptions. For a Retail Revolver product, the “overall indebtedness of the borrower” appears to be the most important factor in predicting LGD, which is strongly correlated with the utilization of borrowers’ available credits and ranked as a top
TABLE 6  Importance of variables for ensemble (bagging) of ten regression trees using all available candidate features (sixteen features selected).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>Utilization of the credit facility</td>
<td>1</td>
</tr>
<tr>
<td>V3</td>
<td>Credit limit of the facility</td>
<td>0.197</td>
</tr>
<tr>
<td>V8</td>
<td>Trend of outstanding balance</td>
<td>0.138</td>
</tr>
<tr>
<td>V4</td>
<td>Joint application indicator</td>
<td>0.102</td>
</tr>
<tr>
<td>V2</td>
<td>Top utilizations of national cards</td>
<td>0.076</td>
</tr>
<tr>
<td>V9</td>
<td>Derived feature on delinquency cycle</td>
<td>0.067</td>
</tr>
<tr>
<td>V10</td>
<td>Derived feature using credit bureau score inform</td>
<td>0.038</td>
</tr>
<tr>
<td>V11</td>
<td>Trend of the card balance</td>
<td>0.030</td>
</tr>
<tr>
<td>V12</td>
<td>Vintage of the account</td>
<td>0.030</td>
</tr>
<tr>
<td>V13</td>
<td>Related loan product count</td>
<td>0.029</td>
</tr>
<tr>
<td>V14</td>
<td>Derived feature using bank’s internal data</td>
<td>0.028</td>
</tr>
<tr>
<td>V15</td>
<td>Derived feature using delinquency cycle inform</td>
<td>0.028</td>
</tr>
<tr>
<td>V16</td>
<td>Vintage of the account</td>
<td>0.027</td>
</tr>
<tr>
<td>V17</td>
<td>Derived feature using spending pattern inform</td>
<td>0.025</td>
</tr>
<tr>
<td>V18</td>
<td>Trend of payment information 1</td>
<td>0.024</td>
</tr>
<tr>
<td>V19</td>
<td>Trend of payment information 2</td>
<td>0.016</td>
</tr>
</tbody>
</table>

predictor; the “ability to pay” is also expected to be strongly predictive, which can be captured by the lower-ranked predictors such as the “joint application indicator” and “net worth”. Hence, the interpretation of the risk factors of a single tree is straightforward.

Now, we fit an ensemble (bagging) model using these seven input features, and the feature importance is presented in Table 5. Compared with the results in Table 4, the ranking of the variables is largely unchanged, even though the numerical values of the importance are different.

Next, we fit a ten-tree ensemble model that selects the sixteen features presented in Table 6. The additional features include delinquency, spending and payment information (V15, V17, V18 and V19). This indicates that the ensemble model can better utilize a large number of available predictors and hence improve the accuracy of the prediction.

Finally, Table 7 shows that the ensemble model improves the prediction accuracy by 8% using the same seven predictors as the regression tree. The performance increase is solely contributed by aggregating multiple models on bootstrapped samples of the training data. The prediction accuracy can be further improved by considering all 499 available inputs and selecting sixteen features using a ten-tree ensemble model.
TABLE 7  Performance comparisons of various methods.

<table>
<thead>
<tr>
<th></th>
<th>RCAP on validation</th>
<th>RCAP on training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression tree</td>
<td>0.364</td>
<td>0.482</td>
</tr>
<tr>
<td>Bagging with same 7 features as candidates</td>
<td>0.390</td>
<td>0.490</td>
</tr>
<tr>
<td>Bagging with all 499 features as candidates</td>
<td>0.403</td>
<td>0.510</td>
</tr>
</tbody>
</table>

TABLE 8  Importance of the variables of the regression tree.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>Percentage below</td>
<td>1</td>
</tr>
<tr>
<td>V2</td>
<td>Effective interest rate</td>
<td>0.647</td>
</tr>
<tr>
<td>V3</td>
<td>Collateral type</td>
<td>0.266</td>
</tr>
</tbody>
</table>

TABLE 9  Importance of variables for ensemble (bagging) of ten regression trees with three candidate features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>Percentage below</td>
<td>1</td>
</tr>
<tr>
<td>V2</td>
<td>Effective interest rate</td>
<td>0.670</td>
</tr>
<tr>
<td>V3</td>
<td>Collateral type</td>
<td>0.174</td>
</tr>
</tbody>
</table>

5.2 Moody’s URD

For Moody’s URD data, the regression tree method selects three features, and “percentage below” is ranked as the most important predictor (see Table 8). The LGD is negatively correlated with the percentage of the debt below, as an obligor’s senior debts always have a lower LGD than junior ones. The “effective interest rate” of the loan or bond is ranked as the second important reflecting the market’s view on the borrowing cost of an obligor, which is closely related to the recovery rate once a loan or bond defaults. This effective interest rate is negatively correlated with interest rates. The third most important variable is the “collateral type”, a clearly important recovery predictor.

Similar to the Retail Revolver data set, we fit an ensemble model using the same three features, and an ensemble of twelve features selected by a ten-tree model considering all candidates. The corresponding variable importance is shown in Tables 9 and 10. Comparing Table 8 with Table 10, the top predictor remains identical, but the lower-ranked predictors are different. This can be explained by the fact that each
TABLE 10  Importance of variables for ensemble (bagging) of ten regression trees with eighteen candidate features (twelve features selected).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
<th>Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>Percentage below</td>
<td>1</td>
</tr>
<tr>
<td>V4</td>
<td>Moody’s industry group</td>
<td>0.531</td>
</tr>
<tr>
<td>V5</td>
<td>Instrument type</td>
<td>0.457</td>
</tr>
<tr>
<td>V2</td>
<td>Effective interest rate</td>
<td>0.42</td>
</tr>
<tr>
<td>V6</td>
<td>Percent above</td>
<td>0.333</td>
</tr>
<tr>
<td>V7</td>
<td>Percentage current</td>
<td>0.172</td>
</tr>
<tr>
<td>V3</td>
<td>Collateral type</td>
<td>0.123</td>
</tr>
<tr>
<td>V8</td>
<td>Downturn indicator</td>
<td>0.078</td>
</tr>
<tr>
<td>V9</td>
<td>Utility indicator</td>
<td>0.073</td>
</tr>
<tr>
<td>V10</td>
<td>Total debt</td>
<td>0.038</td>
</tr>
<tr>
<td>V11</td>
<td>Debt below</td>
<td>0.035</td>
</tr>
<tr>
<td>V12</td>
<td>Collateral ranking</td>
<td>0.027</td>
</tr>
</tbody>
</table>

TABLE 11  Performance comparisons of various methods.

<table>
<thead>
<tr>
<th></th>
<th>RCAP on validation</th>
<th>RCAP on training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Moody’s URD Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Regression Tree</td>
<td>0.605</td>
<td>0.692</td>
</tr>
<tr>
<td>Same three features</td>
<td>0.623</td>
<td>0.626</td>
</tr>
<tr>
<td>All eighteen features</td>
<td>0.631</td>
<td>0.691</td>
</tr>
</tbody>
</table>

Table 11 shows that, compared with a regression tree, model performance on the Moody’s URD data is improved by fitting an ensemble model using the same three features. Further improvement in accuracy can be achieved with the additional features selected.

6 CONCLUSIONS

We used ensemble learning methods to model credit risk. By applying four different ensemble methods on two representative data sets (the Retail Revolver data set and Moody’s Ultimate Recovery Database commercial/capital market data set), our study demonstrates the advantage of ensemble methods. We provided a detailed performance comparison of the ensemble models with the baseline single-regression-tree
model, and an in-depth discussion on model interpretability and the predictive feature selections of the ensemble methods. In addition, we comprehensively discussed the principles underlying performance improvement and, based on these, proposed two new algorithms (ensemble by PLS and bag-boosting) to improve upon the state-of-the-art ensemble (stochastic gradient boost and random forest). The algorithms and implementations are detailed in the online appendixes. We hope the principle and general rules have given readers some insight into how to construct ensemble models tailored to their unique problems. Moreover, as a modeling framework, an ensemble is not limited to using decision or regression trees. Other inference methods can also be combined to increase the performance of predictive models and to tackle complex problems in credit risk modeling.

DECLARATION OF INTEREST

The views expressed in this paper are not necessarily those of Scotiabank, Wells Fargo Bank or any of their affiliates.

ACKNOWLEDGEMENTS

We thank the anonymous reviewer and the Editor for their insightful comments and suggestions and for providing directions for enriching the work presented in this paper.

REFERENCES


Research Paper

Modeling corporate customers’ credit risk considering the ensemble approaches in multiclass classification: evidence from Iranian corporate credits

Parastoo Rafiee Vahid and Abbas Ahmadi

Department of Industrial Engineering and Management Systems, Amirkabir University of Technology, 424 Hafez Avenue, Tehran, Iran; emails: parastoo.rv1@gmail.com, abbas.ahmadi@aut.ac.ir

(Received April 8, 2015; revised February 22, 2016; accepted March 10, 2016)

ABSTRACT

The credit scoring system is one of the most significant credit risk control tools in the banking industry. Usually, the existing credit scoring models classify customers into “good credit” and “bad credit” groups. In this study, a novel model is proposed to classify corporate client accounts into four groups – good credit, past due, overdue and doubtful – according to the definitions of the Central Bank of the Islamic Republic of Iran. This model enables lenders to develop specific policies for credit granting by predicting the solvency and insolvency rates of their corporate clients. For validation, the proposed model, trained by the hybrid approach of a self-organizing map and radial basis function (RBF) neural network, is compared with a single-step, four-class classification model and a model trained by support vector machines. The results show that the proposed model trained by the hybrid approach of a self-organizing map and RBF neural network outperforms the existing methods in terms of its final accuracy with regard to the four classes at the test stage.

Keywords: credit risk; credit scoring model; ensemble approach; artificial neural network; multiclass classification.
1 INTRODUCTION

Credit risk, according to the definition provided by the Basel Committee on Banking Supervision, simply refers to the potential of a bank borrower or counterparty failing to meet its obligations in accordance with agreed terms (Basel Committee on Banking Supervision 1999–2001).

Credit risk management refers to a bank’s risk-adjusted rate of return by maintaining its credit risk exposure within acceptable parameters. In this regard, banks need to manage the credit risk inherent in a portfolio as well as the risk associated with all credit and every single transaction.

Credit scoring is a process by which customers are evaluated quantitatively. These quantitative results predict the ability of customers to meet their repayment obligations in future. In fact, credit assessment models seek to identify those characteristics of (individual or corporate) clients that will affect their likelihood of defaulting on repayments (Hilscher and Wilson 2013).

Although several studies have been conducted to predict customers’ behavior or ratings using statistical techniques and artificial intelligence, the majority of these techniques emphasize dividing customers into “good credit” and “bad credit” groups. They seek solutions to increase the accuracy of the existing models (Abdou et al. 2008; Chuang and Lin 2009). Since the application of these existing models increases the probability of making wrong and/or unfair decisions in credit granting, providing a highly accurate model that is capable of not only identifying the important variables and factors that affect mortgage repayment, but also classifying customers into further categories, seems essential. As a substantial number of facilities are granted to corporate customers in the Iranian banking industry, in this paper the financial ratios of corporate customers are used as input data. We do this in order to provide a credit risk model for four different groups of corporate customers – good credit, past due, overdue and doubtful groups – according to the definitions of the Central Bank of the Islamic Republic of Iran. This enables lenders to develop specific policies for credit granting by predicting the solvency and insolvency rates of their corporate clients. Therefore, the bank’s credit risk decreases, and subsequently its financial resources are used more efficiently.

According to the regulations of the Central Bank of the Islamic Republic of Iran, credit institutions’ receivables are defined as follows.

- Current receivables are claims that are less than two months past their due date.
- Non-current receivables include overdue, past due and doubtful claims, as follows.
  - Overdue receivables are claims that are more than two months and less than six months past their due date or final instalment.
Past due receivables are claims that are more than six months and less than eighteen months past their due date or final instalment.

Doubtful receivables are claims that are more than eighteen months past their due date or final instalment.

The solvent (good credit) group refers to customers without any debts or non-current liability (either in domestic or a foreign currency) to any financial institutions. Other groups include customers with at least one out of three types of non-current debt.

In credit scoring systems that only divide customers into two groups (“good credit” and “bad credit”), all customers with past due, overdue and doubtful debts are considered to be bad credit customers, who will not be approved for any credit in future (Dastoori and Mansouri 2013; Sadatrasoul et al 2014). However, in using the method proposed by this paper, only the doubtful credit group is crossed out. Therefore, by defining notification policies, increasing interest rates or decreasing credit limits, it would be possible to conditionally grant credit to overdue and past due groups.

Implementing a four-class classification technique (a method that discriminates all four groups in a single step) for this issue does not provide the desired accuracy; this is because of the imbalanced data phenomenon that occurs as a result of higher data frequency in the good credit group in comparison with the other groups, as well as the close similarity of the financial statements of two consecutive groups. To overcome these limitations, after collecting financial information about bank customers and determining which features have more influence on model efficiency, the data is divided into sets of 0 and 1, and the multiclass classification problem is broken down into three two-class classification problems. Bagging and random sampling methods are used to solve the problem with inadequate data in all four classes. Then, a radial basis function (RBF) neural network is used for training each of the two-class problems, and the final answer in each class is determined by majority vote for both classes.

With respect to the nature of RBF neural network training, in which supervised and unsupervised learning approaches are employed simultaneously, in order to train this network in the first stage and to determine the parameters of radial functions, a self-organizing neural network has been used.

The rest of the paper is organized as follows. First, there is a literature review of past studies in the field of credit rating and scoring. Second, neural networks, their application in classification and the ensemble approach to multiclass classification are introduced. Then, the proposed method is presented. Finally, the results for the proposed method trained by the RBF neural network, the proposed method trained by the support vector machine (SVM) and the single-step method of the four-class classification are calculated and compared. This is followed by conclusions and recommendations.
2 LITERATURE REVIEW

Credit scoring systems can be divided into three categories (Fensterstock 2003):

(1) judgmental systems;

(2) ranking systems based on statistical techniques;

(3) intelligent systems.

Although statistical techniques such as discriminant analysis and linear regression have been widely used to solve classification problems, the necessity of identifying and analyzing relationships between variables has made them difficult to use in large-scale problems. Other limitations of statistical methods are the assumption of linearity, normality and independent variables in conventional statistical methods. The general regression function consists of a linear probability function, logit function and probit function. The disadvantage of the linear probability function method is that the probability values predicted by this method may be outside of the range (0,1). Although logit and probit models are not limited by normality, and the predicted values are also in the range (0,1), the necessity of transforming the original variables and complicated calculations have made them difficult to use (Chen et al 2009).

The artificial neural network (ANN) has attracted the attention of many scientists and researchers over the last few years and provided amazing results for solving various problems in different areas. An extensive survey of neural network applications in business by Vellido et al (1999) indicates that neural network application in the trade area has more successful outcomes than statistical techniques when solving problems with nonlinear relationships. In fact, the ability of the neural network to identify nonlinear relationships between variables has increased its popularity in the identification and prediction of credit risk compared with parametric (eg, linear discriminant analysis and linear regression analysis) and nonparametric (eg, k-nearest neighbor and decision tree) methods (Chen and Huang 2003).

Glennon et al (2008) utilize a proprietary data set to estimate and validate credit scoring models for credit card holders. The results indicate that current industry best practices can be effective at ranking borrower risk, but they may fall short when it comes to accurately estimating default rates. Modern data-mining techniques play an important role in the field of scientific information and can also be used in developing credit rating models. The computational results produced by Tam and Kiang (1992) confirm that the neural network is the most accurate technique, followed by linear discriminant analysis, logistic regression, decision trees and k-nearest neighbor. In comparison with other techniques, Tam and Kiang (1992) concluded, neural network models are more accurate, robust and adaptive (Oreski et al 2012).
Malhotra and Malhotra (2003) applied a multilayer perceptron neural network to rank twelve American institutions. These results compared favorably with those generated using the discriminant analysis technique in terms of accuracy. Tsai and Wu (2008) demonstrated that neural networks with a back-propagation algorithm are the most popular tools in making financial decisions, due to their unique capability in data classification. Salehi and Mansouri (2011) investigated the performance of logistic regression and neural networks in predicting customers’ credit scores, which confirmed that there exists an equilibrium between these methods. Elter and Yaseen (2010) addressed the importance of financial neural networks in pattern identification as well as time series classification and prediction. Yu et al (2008) emphasized the learning ability of neural networks and proposed a multistage neural network for credit risk analysis. Models evaluating credit applications rely on applicants’ historical repayment performance data, which is only available for approved applicants. Barakova et al (2013) showed that ignoring the rejected applicants significantly affects the forecast accuracy of credit scores, while it has little effect on their discriminatory power.

In recent years, hybrid and ensemble classifiers have been proposed for improving the performance of credit scoring models. Ghodselahi and Amirmadhi (2011) used a hybrid technique consisting of SVMs, neural networks and a decision tree to demonstrate that a hybrid model is more accurate than other scoring techniques. Pacelli and Azzollini (2011) showed the enhanced performance of a hybrid model comprised of neural networks, linear methods and a process analysis of credit risk management.

Finlay (2011) compared the performance of multiple classifiers and found that error trimmed boosting outperformed all other multiple classifiers on UK credit data. Li and Yu (2013) introduced an ensemble model by reviewing different credit assessment techniques. They stated that more accurate classification and prediction can be achieved by using different classification techniques and combining the results.

Sadatrasoul et al (2013) investigated forty-four scientific articles on credit scoring, published by the ScienceDirect database between 2000 and 2012. They also studied the application of neural networks, hybrid and ensemble methods for different companies, small and medium enterprises (SMEs) and individual customers. Their results indicate that the majority of research has been done on individual customers rather than SMEs. In addition, ensemble methods, SVMs and neural networks are among the newest tools used for credit rating.

3 NEURAL NETWORKS

ANNs are based on computational intelligence and attempt to simulate the architecture of neurons in the human brain. So, they are massively connected networks of computational “neurons” and represent parallel-distributed processing structures. Their key
feature is the ability to approximate arbitrary nonlinear functions. ANNs are more effective in machine learning because of the nonlinear nature of decision-making functions. Generally, neural networks are implemented in two steps, as follows.

- **Learning stage**: the network learns general rules from existing data.
- **Test stage**: the learned rules are implemented on new data, and the efficiency of the learning process is studied.

Two well-known ANN topologies are feed-forward and recurrent architectures. In feed-forward neural networks, the entire information flow is forward, and there is no backward flow. The flow starts from the input layer, moves forward layer by layer and ends at the output layer. In this type of network, the output of each stage is independent of the previous stage, and it is used in modeling static systems. Unlike feed-forward neural networks, there are loops in recurrent neural networks, and the information flow can feed back to the same or previous nodes. Thus, this type of network has a memory and is used to model and identify dynamic systems (Hu and Hwang 2002).

### 3.1 RBF networks

RBF networks represent a special category of feed-forward neural network architecture. Researchers developed this connectionist model for estimating the nonlinear behavior of static processes and for function approximation purposes (Hu and Hwang 2002).

The basic RBF network structure consists of an input layer, a single hidden layer with radial activation function (typical transfer functions for hidden functions are Gaussian curves) and an output layer. The standard method used as a radial network learning algorithm is the hybrid method, which is a two-step algorithm (Hu and Hwang 2002; Karry and De Silva 2004).

- The first step involves an unsupervised clustering algorithm, which presents the features of the radial functions used in the network. These features include the centers \( u_i \) and widths \( \sigma_i \) of RBFs. Some of these unsupervised training procedures may be the \( k \)-means-based method, the maximum likelihood estimate-based technique, the standard deviations-based approach and the self-organizing map method. This step is very important for the construction of the RBF network, as an accurate knowledge of \( u_i \) and \( \sigma_i \) has a major impact on the performance of the network.

- The second step includes obtaining the final weights matrix for connections between neurons in the output layer and the hidden layer. In order to update the weight values in the connections between neurons in the hidden layer and
the output layer, a supervised learning method such as gradient descent or least-squares can be used.

If a Gaussian function is used as the radial stimulating function, matrix \( G \) is obtained from the following equations:

\[
G = \{g_{ij}\}, \quad (3.1)
\]

\[
g_{ij} = \exp\left(\frac{-\|x_i - v_j\|^2}{2\sigma_j^2}\right), \quad i = 1, \ldots, n, \quad j = 1, \ldots, r, \quad (3.2)
\]

\[
T = GW, \quad (3.3)
\]

where \( n \) is the number of inputs, \( r \) is the number of hidden nodes and \( T \) is the target value of the training data.

Given that the \( G \) matrix exists only between the output and hidden layers, the weight of the output layer is calculated by \( W = G^{-1}T \) (\( G \) is a square matrix) or \( W = G^+T \) (\( G \) is not a square matrix).

The radial network output is calculated using the following equation (Hu and Hwang 2002):

\[
o_j(x) = \sum_{i=1}^{r} w_{ij} g_i(x), \quad j = 1, \ldots, m, \quad (3.4)
\]

where \( w_{ij} \) is the connection weight between the \( i \)th receptive field unit and the \( j \)th output, and \( g_i \) is the \( i \)th receptive field unit.

With respect to the functional nature of this network (forward and supervised learning) and the number of layers (input, hidden and output), the required approach to training the RBF network is given in Figure 1.

### 3.2 Modular neural network

The modular neural network is a neural network that consists of a number of modules whose functions are all integrated, while each module is a sub-task of a general task in the network. Each module typically functions independently, and the responses of individual modules are combined by an integrating mechanism in order to generate a consolidated response. Like neural networks, modular neural networks can be used to solve various problems such as mapping, function approximation, clustering or associative memory application (Auda and Kamel 1999). Some of the major benefits that make this design approach more attractive than a conventional monolithic global neural network design approach are that it is less complex, more robust, capable of integrating both supervised and unsupervised learning paradigms, more computationally efficient and has a higher learning capacity (Azam 2000).
3.3 Ensemble of neural networks

The modular neural networks can be divided into different types, depending on the strategy used in each of these steps. One of the most common structures of modular
neural network is an ensemble approach based on majority vote. As shown in Figure 2, in this method the final answer is determined from modular results by majority vote (Auda and Kamel 1999).

**4 MODEL DEVELOPMENT**

There are limitations that might affect the development process, such as the amount of existing data, the higher frequency of data in the good credit group versus the other three groups and the similarity of financial ratios between different groups. Thus, in order to overcome these limitations, after collecting customers’ financial information and determining which features have more influence on model efficacy, the multiclass classification problem is converted into three two-class classification problems by dividing the data into sets of 0 and 1. To diminish the effects of inadequate information on all four classes, bagging and random sampling methods are used. Then, the RBF neural network is used for training each of the two-class problems, and the final result in each class is determined by majority vote. The proposed model is given in Figure 3.
4.1 Data preparation

The selection of input (independent) and output (dependent) variables is the first step in the proposed methodology.

(a) Independent variables. Include variables that affect dependent variables; the financial ratios are used as independent variables in this paper. Although the predictive models that use financial ratios have limitations, credit assessment literature highlights the non-substitutability of financial ratios, due to a long history of using them and their concealed information (Chen et al 2009).

(b) Dependent variables. Given that the aim of this paper is to predict banks’ corporate customers’ behavior and categorize them into good credit, past due, overdue and doubtful groups, at first four dependent variables are defined; then, the multiclass classification problem is broken down into three binary classification problems (comprised of 0 and 1 dependent variables), as shown in Figure 3.

Since the data is within different ranges, it may include variables with very small or very large values. To eliminate the possible domination of variables with large values over others, variable values should first be normalized and converted into a certain range of numbers between 0 and 1, or 1 and $-1$. Different methods have been developed for data normalization.

In this paper, the following equation is used for normalizing data and converting it into a new set, in which all data is in the range of 0 to 1:

$$z_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}},$$  \hspace{1cm} (4.1)

where $z_i$ is the normalized value, $x_i$ is the value of each piece of data, $x_{\min}$ stands for the minimum amount of data and $x_{\max}$ refers to the maximum amount of data (Vercellis 2009).

Reducing the number of (predictive variables) features is an important factor in designing a data-mining model. For the most part, the multiplicity of features in data sets with limited observations is not suitable for generating a data-mining model, as the learning algorithm may memorize noise examples, which consequently reduces the accuracy of the final model. Therefore, decreasing the number of features not only reduces the calculation time but also increases the capability of generalizing the data-mining model on the testing data set.

For data reduction, first, a $t$-test is applied to each feature, and the $p$-values obtained from each of them are used as effectiveness criteria to classify credit risk into good and bad groups.

Clustering with a self-organizing map consists of projecting the samples from a high-dimensional data set onto a two-dimensional topological map. Thus, each
ALGORITHM 1 Determination of the features with greatest importance in clustering.

- Define the destiny of each neuron as follows:

$$\gamma_c = \frac{C(c)}{M},$$

(4.2)

where $C(c)$ represents the number of items assigned to the neuron $c$, and $M$ is the number of all neurons on the output layer.

- The importance of variable $j \in \{1, 2, 3, \ldots, d\}$ (feature of database) associated with each neuron $c$ is defined as follows:

$$h_{jc} = \gamma_c \frac{w_{jc}}{\sum_{k=1}^{d} w_{kc}},$$

(4.3)

where $h_{jc}$ is the importance of the $j$th neuron for the $c$th output neuron, $w_{jc}$ is the weight between the $j$th feature and the $c$th output neuron and $\sum_{k=1}^{d} w_{kc}$ is the total weight of all inputs to the $c$th output neuron.

- The importance of the $j$th feature for all neurons in the self-organizing clustering algorithm is calculated using the following equation:

$$\overline{h_j} = \sum_{c=1}^{N} h_{jc}.$$

(4.4)

neuron in the map is activated or not by a subset of data. Using this same feature of the self-organizing neural network, one can identify more important attributes (Benabdeslem and Lebbah 2007). Algorithm 1 is used to determine the features with greatest importance in clustering.

4.2 Dividing data into testing and training groups

Predicting the performance of a model based on its performance with a training data set can be misleading, as the latter is often very optimistic. To predict a model’s effectiveness, another data set should also be used that has not been used for developing the model. Therefore, a data set is divided into two groups: a training group, which includes data only for model creation, and a test (experimental) group for evaluating the model’s performance.

4.3 Training the proposed model

For training this model, first, the data is divided into two sets, 0 and 1, and the multiclass case is broken down into three two-class classification problems. Second, in each defined problem, the bagging method is applied; this is followed by random
ALGORITHM 2 Training the multiclass classification model by using the ensemble approach.

(1) Get training data and related target value.
(2) For classifier \( i \) \((i = 1, 2, 3)\), complete the following:
   (a) select random sets of training data;
   (b) train the RBF network using each random set and obtain the output of two-class classification;
   (c) save the outputs, weights, centers of clusters, radii and accuracy obtained from the training RBF network for each random set;
   (d) when all random sets are checked, calculate the accuracy and error of the given classifier according to majority vote; otherwise, go back to step (b).
(3) Stop when all three classifiers are selected and the output and accuracy of the four classes are obtained; otherwise, remove the error values, specify training data for the next classifier, \( i = i + 1 \), and go back to step (2).

sampling, depending on the size of the case. Third, the RBF neural network is applied for each random set, and the output of each random set is calculated. The final output, which results from the ensemble approach based on majority vote, is then determined for each piece of data (bank corporate customer), according to Algorithm 2.

4.4 Testing the proposed model

According to Figure 3, the following steps are performed for each of the three binary classification problems.

(1) Similar to the training stage, test data is divided into two sets, 0 and 1.
(2) The outputs of the RBF network are calculated for all random sets by using the weights from RBF networks, centers of clusters, radii obtained in the training stage for all random sets and test data.
(3) All outputs obtained by the RBF neural network for testing data are compared with the threshold obtained in the training stage to determine the class of input data, based on the information from each random set.
(4) The final output from the ensemble approach is determined for each piece of data (ie, bank corporate customer), based on the majority vote.
(5) By comparing the classes obtained from the ensemble approach for each piece of data and the target classes, the accuracy and error rates are gained.
TABLE 1 General view of the data set.

<table>
<thead>
<tr>
<th>Class name</th>
<th>Number of samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>111</td>
</tr>
<tr>
<td>Past due</td>
<td>15</td>
</tr>
<tr>
<td>Overdue</td>
<td>10</td>
</tr>
<tr>
<td>Doubtful</td>
<td>34</td>
</tr>
<tr>
<td>Total</td>
<td>170</td>
</tr>
</tbody>
</table>

5 EMPIRICAL ANALYSIS

In this section, the proposed model is trained by financial ratios obtained from corporate client accounts, and the accuracy of the final result is compared with other existing methods.

5.1 Data preparation

This study includes the data for all companies listed on the Tehran Stock Exchange between 2009 and 2013 that were granted short-term loans with a repayment due date of March 21, 2014.

It should be noted that the reason for selecting companies listed on the Tehran Stock Exchange is that their financial statements are audited and disclosed according to the rules and regulations of the Tehran Stock Exchange. The problem with limited accurate and reliable data is obviously solved by this approach, so the developed model will be more precise.

After collecting the information of 200 bank corporate customers, twenty-eight financial ratios were calculated. Finally, given that some financial ratios could not be calculated for some customers, 170 companies were selected to train the model, and their financial ratios were normalized by (4.1). The data frequencies of different groups in the sample are shown in Table 1.

After applying a t-test to each feature, the obtained p-values are considered to be the effectiveness criteria for measuring the ability of each feature to discriminate credit risk into good credit and bad credit groups. Then, the cumulative distribution function is drawn based on the p-values; features with stronger discriminating capabilities will be identified by investigating the obtained results.

As shown in Figure 4, all twenty-eight attributes of p-values are lower than 0.05 (the significance level of the test), indicating that all of the selected attributes have very strong discriminating capabilities. This is because the assumption of a zero-valued mean is rejected in all of them, due to the p-values.

www.risk.net/journal
To determine other important features, the number of output units in self-organizing neural networks should first be determined. Therefore, the optimum number of clusters is determined by sensitivity analysis, based on the problem’s data. As can be seen in Figure 5, eight clusters were selected for this study. Then, the significance of each feature for all neurons in the self-organizing clustering algorithm is calculated.

In terms of discriminating capability and the significance of the feature, and also removing twelve features that are highly correlated with other ratios, the final variables of the model are as follows, in descending order of significance:

(1) working capital;

(2) capital liquidity;

(3) profit margin;
As shown in Figure 6, the three first indexes, namely working capital, capital liquidity and profit margin, have the highest scores in both clustering and discrimination techniques.

Given that the data used for testing should not be employed for network training and that, in dividing data into training and test groups, the data ratio of all four groups should be as similar as possible, the distribution of training and testing data is presented in Table 2.

After determining the number and percentage of each group in the training and testing data sets, members from each of the four groups will be identified, using random selection out of the main data set without replacement.

### 5.2 Train and test data

In this step, 120 training data units are divided into good credit and bad credit groups (past due, overdue and doubtful). Then, the following process will be initiated.

There are different methods for generating random subsets for multiclass classification, such as bagging and different types of boosting techniques (Dudiot and Fridlyand 2003; Optiz and Maclin 1999). As shown in Figure 3, by applying these techniques to all training data, the following results are obtained. In the bagging method with twenty-five replications, the regression error caused by re-implementation declines
FIGURE 6 Comparison between the significance of features based on greatest importance using the clustering technique, and discrimination capability using the \( p \)-value technique.

TABLE 2 Distribution of training and testing data.

<table>
<thead>
<tr>
<th>Class name</th>
<th>% of training data</th>
<th>% of testing data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>67%</td>
<td>62%</td>
</tr>
<tr>
<td>Past due</td>
<td>8%</td>
<td>8%</td>
</tr>
<tr>
<td>Overdue</td>
<td>5%</td>
<td>10%</td>
</tr>
<tr>
<td>Doubtful</td>
<td>20%</td>
<td>20%</td>
</tr>
<tr>
<td>Total percentage</td>
<td>70.58%</td>
<td>29.42%</td>
</tr>
<tr>
<td>Total number of samples</td>
<td>120</td>
<td>50</td>
</tr>
</tbody>
</table>
FIGURE 7 Comparison between the errors of bagging and boosting methods.

- The training error decreases toward 0, while this error is stable in the boosting method. Therefore, the bagging method is used here to generate information and to combine the results of several primitive classifications.

- According to the model presented in Figure 3, the functional nature of the RBF neural network and the number of its layers (ie, input, hidden and output), an RBF network that can satisfy the requirements of the first classifier problem must have the following characteristics and structure.

  1. Since the problem’s inputs are comprised of sixteen features, sixteen neurons are considered in the input layer.
  2. There is just one hidden layer.
  3. There are eight centers and neurons in the hidden layer, which are obtained using a self-organizing neural network.
  4. Given the type of problem (two-class classification), one neuron is considered in the output layer.
  5. The radii of radial function, after data clustering, are obtained from standard deviation of the points inside each cluster.
  6. The Gaussian and linear functions are used in the hidden and output layers, respectively.
TABLE 3 Characteristics of RBF network for each of the classifiers.

<table>
<thead>
<tr>
<th>Name of classifier</th>
<th>Number of random sets</th>
<th>Number of layers</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classifier 1</td>
<td>25</td>
<td>16</td>
<td>0.661</td>
</tr>
<tr>
<td>Classifier 2</td>
<td>15</td>
<td>4</td>
<td>0.357</td>
</tr>
<tr>
<td>Classifier 3</td>
<td>7</td>
<td>2</td>
<td>0.602</td>
</tr>
</tbody>
</table>

(7) The weights of the output layer are obtained using a supervised “gradient descent”-based learning algorithm.

(8) The number of training periods is equal to 200, or the error rate is lower than $1 \times 10^{-8}$.

(9) For each of the twenty-five random sets, the network’s centers (using a self-organizing neural network), radii and output are calculated separately.

(10) After training all random sets, the final output is determined by majority vote.

To train the second classifier (discriminating the doubtful group from the overdue and past due groups), error data from the previous step is initially deleted. Due to the reduced data number, fifteen random data sets are selected from the remaining data. In the clustering stage with a self-organizing neural network, the number of clusters is considered to be four; therefore, centers and radii are calculated based on four neurons in the middle layer of the RBF neural network. After applying these changes, steps (1) to (10) are repeated, output sets are determined for each of the random sets and the final output is determined by majority vote.

Finally, to train the third classifier (discriminating overdue from the past due group), seven random sets of the remaining data are selected after eliminating error data from the previous stage, the number of middle layer neurons is set to two and all the previous steps are repeated. The required changes to train each of the three classifiers are shown in Table 3.

As is shown in Figure 8, the threshold number for each of the three classifiers is determined by sensitivity analysis.

In the testing stage, similar to the training stage, fifty testing data units are divided into good credit and bad credit groups (past due, overdue and doubtful) for the first classifier. Then, by using the weights, cluster centers, radii and threshold determined for twenty-five random sets in the training stage, twenty-five output sets are calculated from testing data. The final output and the model’s accuracy are obtained through majority vote. To test the other two classifiers, first errors from the previous stage are eliminated, and then all the previous steps are repeated.
5.3 Implementation of the proposed model using SVMs

As mentioned, SVMs are generally binary and linear classifiers, which are utilized as multiclass and nonlinear classifiers by using Kernel functions (Bellotti and Crook 2009; Huang et al 2007). In the proposed model, SVMs have been employed instead of the self-organizing map and RBF neural networks for clustering and classification. In addition, all steps of the algorithm, including three-stage classifiers, removing error data in the training stage, random sampling and testing, are exactly replicated. The final accuracy level obtained from implementing the model proposed by SVMs is presented in Table 5.
TABLE 4  Accuracy of the model trained by the hybrid approach of RBF and SOM neural networks.

(a) Train

<table>
<thead>
<tr>
<th></th>
<th>Classifier 1</th>
<th>Classifier 2</th>
<th>Classifier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>87.5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past due</td>
<td></td>
<td>77.7%</td>
<td></td>
</tr>
<tr>
<td>Overdue</td>
<td>70%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Doubtful</td>
<td></td>
<td>80%</td>
<td>100%</td>
</tr>
</tbody>
</table>

(b) Test

<table>
<thead>
<tr>
<th></th>
<th>Classifier 1</th>
<th>Classifier 2</th>
<th>Classifier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>83.9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past due</td>
<td></td>
<td>88%</td>
<td></td>
</tr>
<tr>
<td>Overdue</td>
<td>100%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Doubtful</td>
<td></td>
<td>80%</td>
<td>75%</td>
</tr>
</tbody>
</table>

5.4 Analysis of the results

In order to assess the validity of the proposed model, the model trained by the hybrid approach of self-organizing and RBF neural networks, the model trained by SVMs and the single-step, four-class classification model trained by the hybrid approach of self-organizing and RBF neural networks are compared. The results are summarized in Table 6.

Given that, in the three-stage classifiers training in models A and B, error data is removed at each stage, and training in the next stage is done using correctly classified data and by comparing the results obtained from the training data, it can be found that in the training stage by SVMs, less data is recognized as error data and removed. A comparison between the results obtained from the implementation of the models using training and testing data reveals that the model trained by SVMs suffers from over-fitting, and its results are not reliable. The over-fitting phenomenon happens when a model’s training errors are less than the other model, while the inverse is true in the testing stage, which is the final criterion for a model’s performance assessment. This problem may be due to the noise and complicated nonlinear relationships between the input variables in the model, or because of the small amount of training data. In contrast, the hybrid model of self-organizing and RBF neural networks is an accurate model with reliable results.
TABLE 5  Accuracy of the model trained by SVMs.

(a) Train

<table>
<thead>
<tr>
<th></th>
<th>Classifier 1</th>
<th>Classifier 2</th>
<th>Classifier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>96.25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past due</td>
<td></td>
<td>78.3%</td>
<td></td>
</tr>
<tr>
<td>Overdue</td>
<td>92.3%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Doubtful</td>
<td></td>
<td>90%</td>
<td>100%</td>
</tr>
</tbody>
</table>

(b) Test

<table>
<thead>
<tr>
<th></th>
<th>Classifier 1</th>
<th>Classifier 2</th>
<th>Classifier 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good credit</td>
<td>87%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Past due</td>
<td></td>
<td>75%</td>
<td></td>
</tr>
<tr>
<td>Overdue</td>
<td>84%</td>
<td></td>
<td>100%</td>
</tr>
<tr>
<td>Doubtful</td>
<td></td>
<td>87.5%</td>
<td>67%</td>
</tr>
</tbody>
</table>

Due to the imbalanced data phenomenon that occurs as a result of higher data frequencies in the good reputation group in comparison with the other groups, and because of the close similarity of the financial statements between two consecutive groups, the single-step, four-class classification model trained by a hybrid approach of self-organizing and RBF neural networks does not provide the desired accuracy.

6 CONCLUSION

In this research, a model was proposed for four-class classification of credit risk of a bank’s corporate customers by an ensemble approach. This paper mainly aims to offer a model for differentiating the middle group(s) between good credit and bad credit. The study recommends a flexible model to reduce banks’ credit risks while providing them with a tool to grant credit to a broader range of applicants. As the model is proposed for multiclass prediction and classification, it is valid in the Iranian credit system, regardless of this market’s isolation from the global market in recent years or imposed economic sanctions. It is also applicable to credit systems in other countries with similar groupings, ie, creditworthy, non-creditworthy and in-between groups. Given the complicated nature of nonlinear relationships between the model’s input variables and imbalanced data phenomenon in this study, the application of common multiclass classification techniques by which the information is divided into the four desired groups in a single stage will not provide suitable accuracy. Therefore, by differentiating information into binary groups, we have tried to eliminate this
TABLE 6  Comparison between final accuracy of three models in test step.

<table>
<thead>
<tr>
<th>Model A</th>
<th>Final accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class name</td>
<td></td>
</tr>
<tr>
<td>Good credit</td>
<td>84%</td>
</tr>
<tr>
<td>Past due</td>
<td>80%</td>
</tr>
<tr>
<td>Overdue</td>
<td>75%</td>
</tr>
<tr>
<td>Doubtful</td>
<td>80%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model B</th>
<th>Final accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class name</td>
<td></td>
</tr>
<tr>
<td>Good credit</td>
<td>87%</td>
</tr>
<tr>
<td>Past due</td>
<td>60%</td>
</tr>
<tr>
<td>Overdue</td>
<td>50%</td>
</tr>
<tr>
<td>Doubtful</td>
<td>70%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model C</th>
<th>Final accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class name</td>
<td></td>
</tr>
<tr>
<td>Good credit</td>
<td>60%</td>
</tr>
<tr>
<td>Past due</td>
<td>30%</td>
</tr>
<tr>
<td>Overdue</td>
<td>33%</td>
</tr>
<tr>
<td>Doubtful</td>
<td>17%</td>
</tr>
</tbody>
</table>


problem. In addition, to eliminate the data limitation problem in some groups (past due and overdue), we have attempted to improve the classification accuracy, as much as possible, by applying bagging and random sampling methods, training every random set separately and, finally, combining the results from random sets based on majority vote. To train the proposed model, a RBF neural network was employed. The training RBF neural network is a kind of hybrid method in which the parameters (centers and radii) are first obtained using an unsupervised self-organizing neural network, and then the weight of the output layers is obtained using a supervised algorithm. In order to assess the validity of the proposed model, the model was once again trained by SVMs, and then the two models were compared. By comparing the results,
it was determined that the model trained by SVMs suffered from the over-fitting phenomenon, while the results obtained by the proposed hybrid model of RBF and a self-organizing neural network were appropriately reliable. In future research, more detailed studies of corporate customers may determine what factors cause a customer to move from one class to another. Also, using market value information and adding market value ratio to other financial ratios may lead to more accurate results and, thus, increase the accuracy of the model.

DECLARATION OF INTEREST

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

REFERENCES


