JOURNAL OF APPLIED STATISTICS
2022, VOL. 49, NO. 10, 2700–2716
https://doi.org/10.1080/02664763.2021.1910936

Taylor & Francis
Taylor & Francis Group

APPLICATION NOTE

Assessing the discriminatory power of loss given default models

Hannes Kazianka[a], Anna Morgenbesser[b] and Thomas Nowak[b]

[a]Department of Statistics, University of Klagenfurt, Klagenfurt, Austria; [b]Oesterreichische Nationalbank, Vienna, Austria

ABSTRACT
For banks using the Advanced Internal Ratings-Based Approach in accordance with Basel III requirements, the amount of required regulatory capital relies on the banks' estimates of the probability of default, the loss given default and the conversion factor for their credit risk portfolio. Therefore, for both model development and validation, assessing the models' predictive and discriminatory abilities is of key importance in order to ensure an adequate quantification of risk. This paper compares different measures of discriminatory power suitable for multi-class target variables such as in loss given default (LGD) models, which are currently used among banks and supervisory authorities. This analysis highlights the disadvantages of using measures that solely rely on pairwise comparisons when applied in a multi-class setting. Thus, for multi-class classification problems, we suggest using a generalisation of the well-known area under the receiver operating characteristic (ROC) curve known as the volume under the ROC surface (VUS). Furthermore, we present the `R`-package `VUROCS`, which allows for a time-efficient computation of the VUS as well as associated (co)variance estimates and illustrate its usage based on real-world loss data and validation principles.

ARTICLE HISTORY
Received 10 April 2020
Accepted 27 March 2021

KEYWORDS
Loss given default; volume under surface; discriminatory power; receiver operating characteristic

2010 MATHEMATICS SUBJECT CLASSIFICATION
62P05

1. Introduction

With the implementation of the Basel II capital adequacy rules [1], banks have been given the possibility to use own internal risk models to estimate the probability of default (PD), loss given default (LGD) and conversion factors (CF) for their credit risk portfolio. After the approval by national supervisory authorities, banks use these models to estimate their obligors' or facilities' PD, LGD and CF for the purpose of own funds requirements calculation. Thus, it is of crucial importance that the risk models generate valid and reliable estimates of the true (yet unknown) risk parameters. To that end, the EU's Capital Requirements Regulation (CRR; Regulation (EU) No 575/2013), among others, requires banks to implement a regular cycle of model validation to assess the capabilities of these risk models. The European Central Bank (ECB) for example, recognises the key role of adequate validation procedures by imposing requirements for the bank's internal validation routines [6]

CONTACT Hannes Kazianka ✉ hannes.kazianka@aau.at 🖃 Department of Statistics, University of Klagenfurt, Universitätsstraße 65–67, Klagenfurt, 9020, Austria

as well as systematically collecting information from the banks' internal validation units [7] to allow for cross-comparisons among the supervised entities.

The purpose of measuring discriminatory power of credit risk models is to assess how good a model can differentiate between obligors or facilities with high and low risk, which in turn is a key ability for a model's use in risk management. In this context, the ECB [6] expects heterogeneity of obligors or facilities between obligor/facility rating grades and homogeneity within grades. In the case of models for the risk parameter PD, we are dealing with a binary classification problem, where models need to predict defaults and non-defaults. For these models, a number of reliable measures (e.g. the accuracy ratio) have become best practice among banks and in the literature (for an overview see, for example, [4,14,17]). However, as pointed out by both [15,16], the measurement of discriminatory power for LGD models used by banks and supervisory authorities is still at an early stage. This may be caused by the increased complexity when evaluating a model's discriminatory power since the outcome of LGD and CF models usually is continuous. Thus, for the purpose of discriminatory power assessment, the predicted and realised LGDs are usually mapped to a discretised LGD scale and one is faced with a multi-category classification problem.

To the best of the authors' knowledge, no common understanding of well-suited measures of discriminatory power for LGD models has yet been established among banks and supervisory authorities. At the same time, the validity of the LGD parameter is of high importance as the Basel risk-weight functions (see [1]) and, ultimately, the banks' own funds requirements are highly sensitive to changes in the parameter LGD. In this context, recent discussions among supervisory authorities raised concerns about the consistency of the interpretation of regulatory requirements across banks and jurisdictions, causing unwarranted variability in capital requirements for credit risk. In particular, the European Banking Authority (EBA) performed several studies benchmarking the risk parameter estimates across different EU countries and identified significant differences in modelling practices of risk parameter estimates (for the latest available benchmarking exercise, see [5]). As a consequence, the ECB launched the Targeted Review of Internal Models (TRIM) with the aim of harmonising supervisory practices within the Single Supervisory Mechanism (SSM). Moreover, the ECB established a common reporting standard of validation results of internal models with the aim of strengthening model validation. In conclusion, comprehensive methods to validate the calibration and discriminatory power of LGD models are an essential tool to assess the reliability of parameter estimates and may help in further strengthening their credibility.

The remainder of this paper is organised as follows: Section 2 first presents an overview of discriminatory power measures commonly used in practice and shows the drawbacks of these measures in assessing the discriminatory power of LGD models. Following the work of [12,18], Section 3 generalises the analysis of receiver operating characteristic (ROC) curves to allow for the analysis of multi-class classification problems. This leads to the introduction of the volume under the ROC surface (VUS) as a suitable discriminatory power measure in a multi-class setting. To increase the accessibility of the numerical routines to compute the VUS and its (co)variance, we present the R-package VUROCS available from the Comprehensive R Archive Network (CRAN) at https://CRAN.R-project.org/package=VUROCS. It comes along with time-efficient implementations of classical ordinal measures of association such as Somers' D, Kruskal's Γ

or Kendall's τ_b and τ_c. Finally, Section 4 illustrates the advantages of VUS compared to commonly used discriminatory power measures in LGD models on the basis of a hypothetical example, real world loss data of a large commercial bank and validation principles. Section 5 summarises the main conclusions and closes the paper.

2. Overview of existing discriminatory power measures

Among banks, a number of measures have been proposed to test the discriminatory power of LGD models. Supervisory authorities have implicitly accepted some of these measures when approving the banks' validation concepts. Moreover, they are themselves using bivariate measures, e.g. in TRIM inspection tools and the validation reporting [7]. This section provides a brief overview of discriminatory power measures which are commonly used for validating the discriminatory power of LGD models. The terminology used throughout the paper is as follows: $n \in \mathbb{N}$ is the total number of observations (in general, customer facilities) and n_k the number of classified realised LGD values $(y_1, \ldots, y_n) \in \mathcal{Y}^n$ which fall into LGD grade $\bar{y}_k \in \mathcal{Y}$ for $k \in \{1, \ldots, r\}$. Thereby, a linear order relation $\leq_\mathcal{Y}$ is defined on the elements of $\mathcal{Y}$. Furthermore, $f : \mathcal{X} \to \mathbb{R}$ is a ranking function that imposes an ordering on the explanatory variables $(x_1, \ldots, x_n) \in \mathcal{X}^n$ which can ultimately be converted to predicted LGD grades via a monotonous step-function h via $h(f(\cdot)) = h_f : \mathcal{X} \to \mathcal{Y}$.

2.1. Cumulative LGD accuracy profile

Similar to the ROC used for assessing the discriminatory power of PD models, [15] suggested to construct a power curve for LGD models on grade level, the so-called cumulative LGD accuracy profile. To measure the ordinal ranking power, the cumulative LGD accuracy ratio (CLAR) is then calculated based on the area under this power curve. Starting at the origin $(0, 0)$, the curve connects additional r points, which are derived as follows. The coordinates on the abscissa for the points on the power curve are determined by the cumulative number of observations in each predicted LGD grade (order from high to low) divided by the total number of observations n:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(\bar{y}_k \leq_\mathcal{Y} h_f(x_i)), \quad k = r, r-1, \ldots, 1.$$

Here, $\mathbb{1}(\cdot)$ denotes the indicator function. Furthermore, on the ordinate, the cumulative percentage of correctly assigned observations is displayed:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}((\bar{y}_k \leq_\mathcal{Y} h_f(x_i)) \wedge (\bar{y}_k \leq_\mathcal{Y} y_i)), \quad k = r, r-1, \ldots, 1.$$

The CLAR measure is then defined as twice the area under the curve. CLAR takes values between 0 and 1, where a value close to 1 suggests that the model is able to discriminate well between high and low risk facilities, while values near 0 indicate a poor discriminatory power of the LGD model.

The following example shows one of the limitations of CLAR for discriminatory power assessment. Consider an LGD model with two grades $\bar{y}_1$ and $\bar{y}_2$. Assume that there are two

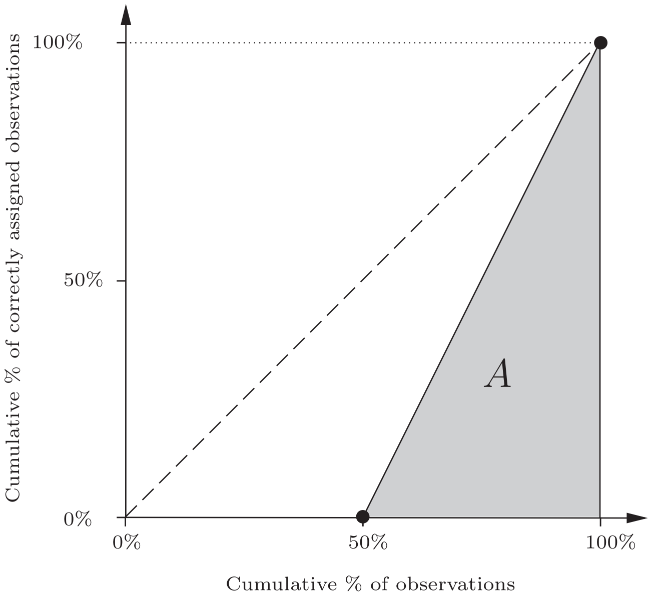

Figure 1. Cumulative LGD accuracy profile for the example $(y_1, h_f(x_1)) = (\bar{y}_1, \bar{y}_2)$ and $(y_2, h_f(x_2)) = (\bar{y}_2, \bar{y}_1)$ with $CLAR = 2 \cdot A$

observations with perfectly reverse order between the realisation and the prediction, i.e. $(y_1, h_f(x_1)) = (\bar{y}_1, \bar{y}_2)$ and $(y_2, h_f(x_2)) = (\bar{y}_2, \bar{y}_1)$. We would expect that a proper measure of discriminatory power would return the worst possible value for this constellation. However, CLAR does not result in a value of zero or close to zero. CLAR for this example is 0.5, hence, covering up the fact that this model cannot discriminate between borrowers with high LGD and low LGD at all. Figure 1 presents a graphical illustration of the power curve for this example. The fact that CLAR returns data-dependent positive values even though the LGD model would fail to rank even one observation correctly renders CLAR non-comparable across different data sets. However, this could be easily corrected using the following adjustment:

$$\mathrm{CLAR}^{adj} = \frac{\mathrm{CLAR} - c}{1 - c},$$

where c denotes the value of CLAR for the worst possible ranking of the considered data (in the above example $c = 0.5$). Hence, CLAR^{adj} will return a value of zero if the two ordinal rankings of predicted and realised LGD are in reverse order.

2.2. Ordinal measures of association

Apart from CLAR, a number of discriminatory power measures already applied for LGD models exist which rely on pairwise comparisons to assess the correctness of ordinal rankings. To that end, ordinal measures of association (or 'measures of concordance'; see Definition 5.1.7 in [13]) are popular among practitioners and currently also widely accepted by supervisory authorities. These statistics are used to assess the rank correlation of two orderings imposed by two different ranking functions. Popular measures of ordinal association are Somers' D, Kruskal's Γ, Kendall's τ_b, Kendall's τ_c and Spearman's ρ. In the following, we will limit our focus on the two commonly used measures Somers' D as well as Kendall's τ_b. In line with the definitions of [2], let a_{ij} be the reported frequency of the ith row and jth column of the $\mathcal{Y} \times \mathcal{Y}$ or $\mathcal{Y} \times f(\mathcal{X})$ contingency table. Define $r_i = \sum_j a_{ij}$ and $c_j = \sum_i a_{ij}$ for the ith row sum and the jth column sum, respectively. Again, $n = \sum_i \sum_j a_{ij}$ is the total number of facilities being ranked. To simplify notation, we introduce the following variables

$$A_{ij} = \sum_{k<i} \sum_{l<j} a_{kl} + \sum_{k>i} \sum_{l>j} a_{kl}, \quad D_{ij} = \sum_{k>i} \sum_{l<j} a_{kl} + \sum_{k<i} \sum_{l>j} a_{kl}$$

$$P = \sum_i \sum_j a_{ij} A_{ij}, \quad Q = \sum_i \sum_j a_{ij} D_{ij}.$$

In line with [2], the measures Somers' D and Kendall's τ_b are then defined as:

$$\text{Somers}'\text{D} = \frac{P - Q}{n^2 - \sum_i r_i^2}, \quad \tau_b = \frac{P - Q}{\sqrt{(n^2 - \sum_i r_i^2)(n^2 - \sum_j c_j^2)}}.$$

Both measures range between -1 and 1, where -1 indicates a perfect negative association and 1 a perfect positive association. In order to allow for comparison with CLAR^{adj}, which attains values between 0 and 1, we will rescale Somers' D and Kendall's τ_b to the same range.

$$\text{Somers}'\text{D}^{rescaled} = \frac{1 + \text{Somers}'\text{D}}{2} \quad \tau_b^{rescaled} = \frac{1 + \tau_b}{2}$$

Note that in the dichotomous case, Somers' D is equivalent to the well-known accuracy ratio, being linearly related to the area under the ROC curve, which is commonly used to assess the discriminatory power of PD models.

2.3. Measures based on the area under the Kendall curve

To graphically explore the bivariate dependence structure of an arbitrary pair of random variables (U, V), [8] propose to plot the Kendall distribution function $K(t) = \mathbb{P}\{H(U, V) < t\}$ against $t - t\log(t)$ with $t \in [0, 1]$. Here, $H(u, v) = \mathbb{P}\{U < u, V < v\}$ is the joint distribution function of (U, V). They called the resulting plot the K-plot and the curve $(K(t), t - t\log(t))$ the Kendall curve. Inspired by the area under the ROC curve, [19] recently proposed the area under the Kendall curve (AUK) as a measure:

$$\begin{aligned} AUK &= 1 - \int_0^1 K(t)\,\mathrm{d}(t - t\log(t)) = 1 + \int_0^1 K(t)\log(t)\,\mathrm{d}t \\ &= \mathbb{E}\left\{H(U, V) - H(U, V)\log\left(H(U, V)\right)\right\}. \end{aligned}$$

See also [20] for more details on the AUK's properties. Note that $K(t)$ is not only defining AUK but also Kendall's τ via $\tau = 4\int_0^1 tdK(t) - 1$ (see Equation (5.1.10) in [13]).

Considering $H_0 = H$, $H_1(u, v) = \mathbb{P}\{U \geq u, V < v\}$, $H_2(u, v) = \mathbb{P}\{U < u, V \geq v\}$ and $H_3(u, v) = \mathbb{P}\{U \geq u, V \geq v\}$, [20] defined $K_i(t) = \mathbb{P}\{H_i(U, V) < t\}$ and $AUK_i = 1 -$

$\int_0^1 K_i(t)\log(t)\,\mathrm{d}t,\ i=0,1,2,3$ to extend AUK to a measure of dependence. We take over these definitions and define further

$$AUK_{tot} = \frac{2}{3}\,(AUK_0 + AUK_3 - AUK_1 - AUK_2)$$

that we propose as a novel ordinal measure of association. In fact, contrary to AUK, AUK_{tot} satisfies the defining properties of Definition 5.1.7 in [13]. This follows from the properties of AUK_i implied by Proposition 2 in [20], the symmetry of $K_i(t)$ in U and V, the equicontinuity of copulas, and by inferring that AUK_0 and AUK_3 as well as $-AUK_1$ and $-AUK_2$ are increasing functions in the concordance ordering of copulas (see Definition 2.8.1 in [13]).

When using AUK and AUK_{tot} for the purpose of measuring discriminatory power of LGD models, H is replaced by the empirical distribution function of the observations in the $\mathcal{Y}\times\mathcal{Y}$ or $\mathcal{Y}\times f(\mathcal{X})$ space and AUK_i is estimated empirically via the relation $AUK_i = \mathbb{E}\{H_i(U,V) - H_i(U,V)\log(H_i(U,V))\},\ i=0,1,2,3.$

3. The volume under the ROC surface and its efficient computation

This section first provides a formal definition of the measure VUS, which is a generalisation of the area under the ROC curve for multi-category classification problems. We will then present an efficient algorithm as defined by [18] to calculate VUS as well as its (co)variance. Subsequently, we provide an implementation of this algorithm in the R-package VUROCS.

3.1. Volume under the ROC surface

As outlined above, the area under the ROC curve (or equivalent measures such as the accuracy ratio) is frequently used to assess the discriminatory power of PD models. However, this measure is only applicable for dichotomous classification problems, where there are only two outcomes, e.g. default or non-default. For LGD models, where one is usually faced with more than two grades for the prediction of LGD, the standard measures based on ROC curves are not applicable. To assess the discriminatory power of multi-class classification problems, [9] derive their measure based on a generalisation of the area under the ROC curve by averaging the results from pairwise comparisons. In contrast to this approach, [3] extended the ROC methodology for three-class problems by considering three-dimensional ROC surfaces. Nakas and Yiannoutsos [12] extend this approach to an r-dimensional ROC surface for objects coming from r different ordered categories called the volume under the ROC surface (VUS). They also propose non-parametric estimators for the variance of VUS using results for U-statistics. Nakas and Yiannoutsos [12] conclude that the computational burden becomes intractable for big data sets and high dimension, which might cause its diffident practical application. Still, the statistic VUS is already in use to assess the discriminatory power of multi-category classification problems encountered, among others, in medical decision problems [11]. However, to the best of the authors' knowledge, the application of VUS in the validation of credit risk models has not yet been analysed in the literature or among practitioners. In order to overcome the computational issues in computing VUS, [18] present efficient algorithms for calculating the VUS and estimators of its variance and covariance. These algorithms scale well with respect to the number of grades and the size of data sets. The time complexity of the algorithm for estimating the variance and covariance of VUS is $\mathcal{O}(2^r n^2)$ and $\mathcal{O}(2^r n^4)$, respectively.

The following describes the generalisation of the area under the ROC curve for multi-category classification problems. Using the notation of Section 2 and [18], the VUS is given by

$$\hat{U}(f,D) = \frac{1}{\prod_{k=1}^{r} n_k} \sum_{j_1,\dots,j_r \in \{1,\dots,n\}: y_{j_1} < \dots < y_{j_r}} \mathbb{1}(f(x_{j_1}) < \dots < f(x_{j_r})),$$

where j_k runs over all facilities which are assigned to grade $\bar{y}_k$ and $D = \{(x_1,y_1),\dots,(x_n,y_n)\}$ with $D \subseteq \mathcal{X} \times \mathcal{Y}$. In contrast to the ordinal measures of association measures presented in Section 2.2, the VUS considers the ordering of r-tuples instead of relying on pairwise comparison. Evidently, in case of a random classifier, $\mathbb{E}\{\hat{U}(f,D)\} = 1/r!$. Furthermore, [12] derived an expression of the variance of VUS, which allows for statistical inference. In the r-dimensional generalisation, [18] first define the set of splits of $\mathcal{Y}$ as $\Upsilon := \{(\mathcal{Z}_1,\mathcal{Z}_2) | \mathcal{Z}_1 \cup \mathcal{Z}_2 = \mathcal{Y} \wedge \mathcal{Z}_1 \cap \mathcal{Z}_2 = \varnothing\}$ and introduce the expression

$$q_v(f,\mathcal{Z}_1,\mathcal{Z}_2) := \mathbb{P}\{(f(X_1) < \dots < f(X_r)) \wedge (f(X_1') < \dots < f(X_r'))\},$$

The explicit formula for the variance of the VUS is:

$$\sigma^2_{\hat{U}}(f,D) := \frac{1}{\Pi_{k=1}^{r} n_k} \sum_{(\mathcal{Z}_1,\mathcal{Z}_2)\in\Upsilon} \left(\Pi_{\bar{y}_l \in \mathcal{Z}_2}(n_l - 1)\right)(q_v(f,\mathcal{Z}_1,\mathcal{Z}_2) - \hat{U}(f,D)^2),$$

where $q_v(f,\mathcal{Z}_1,\mathcal{Z}_2)$ and $\hat{U}(f,D)$ can be estimated for fixed D. The covariance for the volumes of two ranking functions f_1 and f_2 can be derived in a similar manner,

$$\begin{aligned}&\mathrm{Cov}_{\hat{U}}(f_1,f_2,D)\\ &:= \frac{1}{\Pi_{k=1}^{r} n_k} \sum_{(\mathcal{Z}_1,\mathcal{Z}_2)\in\Upsilon} \left(\Pi_{\bar{y}_l \in \mathcal{Z}_2}(n_l - 1)\right)(q_c(f_1,f_2,\mathcal{Z}_1,\mathcal{Z}_2) - \hat{U}(f_1,D)\hat{U}(f_2,D)),\end{aligned}$$

where $q_c(f_1,f_2,\mathcal{Z}_1,\mathcal{Z}_2) := \mathbb{P}\{(f_1(X_1) < \dots < f_1(X_r)) \wedge (f_2(X_1') < \dots < f_2(X_r'))\}$.

As the algorithms proposed by [18] for calculating the variance and covariance of VUS show some minor errors, we present corrected versions of the algorithms which we use to compute `VUSvar()` and `VUScov()` from the R-package `VUROCS` (see Section 3.2). For efficient calculation of the (co)variance of VUS, [18] use (compressed) directed layered graphs $G = (V,E)$ with $r+2$ layers for every split of $\mathcal{Y}$ into two disjoint sets $\mathcal{Z}_1$ and $\mathcal{Z}_2$, where V is the set of nodes and E is the set of edges of the graph. Furthermore, the first layer V_0 and the last layer V_{r+1} include a start node v_s and an end node v_e. Finally, $\Gamma(v)$ counts the number of distinct paths starting from node v_s to node v, with the share of $\Gamma(v_e)$

in all possible paths being an estimator for q_v and q_c, respectively. For further details on notation and rationale behind the algorithm, see [18].

```
Input: data set $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, ranking function $f$;
Sort $D$ according to $f$;
Compute $\hat{U}(f, D)$;
$\sigma^2_{\hat{U}}(f, D) = 0$;
for $(\mathcal{Z}_1, \mathcal{Z}_2) \in \Upsilon$ do
    Extend $\mathcal{Z}_1$ with $\bar{y}_0$ and $\bar{y}_{r+1}$;
    Extend $D$ with a dummy head $\tilde{v}_s = (x_0, y_0)$ and dummy tail $\tilde{v}_e = (x_{n+1}, y_{n+1})$;
    $y_0 = \bar{y}_0,\ y_{n+1} = \bar{y}_{r+1},\ p = 0$;
    for $k = 0$ to $r + 1$ do
        if $\bar{y}_k \in \mathcal{Z}_1$ then
            $\tilde{V}_p = \{(x_i, y_i) | y_i = \bar{y}_k\}$
            $p = p + 1$
        end
    end
    for $p = 0$ to $|\mathcal{Z}_1| - 2$ do
        connect nodes of consecutive layers $\tilde{V}_p$ and $\tilde{V}_{p+1}$ and set weights to zero;
        $\bar{y}_k =$ category associated with layer $\tilde{V}_p$;
        $\bar{y}_l =$ category associated with layer $\tilde{V}_{p+1}$;
        if $l = k + 1$ then
            $\forall (x_i, y_i), (x_j, y_j) : y_i = \bar{y}_k \wedge y_j = \bar{y}_l \wedge f(x_i) < f(x_j) \implies w((x_i, y_i), (x_j, y_j)) = 1$;
        end
        else
            for $y_i = \bar{y}_k$ do
                $w_h = 0$ for $h = 1, \ldots, l - k - 1$
                for $j = i + 1$ to $n + 1$ do
                    $y_j$ corresponds to category $\bar{y}_m$;
                    if $m = k + 1$ then
                        $w_1 = w_1 + 1$;
                    end
                    if $k + 1 < m < l$ then
                        $w_{m-k} = w_{m-k} + w_{m-k-1}$
                    end
                    if $m = l$ then
                        $w((x_i, y_i), (x_j, y_j)) = w_{l-k-1}$
                    end
                end
            end
        end
    end
    $\tilde{\Gamma}(\tilde{v}_s) = 1; \forall\, v_a \in \tilde{V}_k, k \geq 1 : \tilde{\Gamma}(v_a) = 0$;
    for $k = 0$ to $|\mathcal{Z}_1| - 2$ do
        for $v_a \in \tilde{V}_k$ do
            for $v_b \in \tilde{V}_{k+1}$ do
                $\tilde{\Gamma}(v_b) = \tilde{\Gamma}(v_b) + w^2(v_a, v_b)\tilde{\Gamma}(v_a)$
            end
        end
    end
    $\sigma^2_{\hat{U}}(f, D) = \sigma^2_{\hat{U}}(f, D) + (\Pi_{\bar{y}_l \in \mathcal{Z}_2}(n_l - 1))(\frac{\tilde{\Gamma}(\tilde{v}_e)}{\underline{n}} - \hat{U}(f, D)^2)$
    where $\underline{n} = \prod_{k=1}^{r} n_k \prod_{\bar{y}_l \in \mathcal{Z}_2} n_l$ is the number of all theoretically possible paths from $\tilde{v}_s$ to $\tilde{v}_e$ if all nodes
      between subsequent layers were connected
end
$\sigma^2_{\hat{U}}(f, D) = \frac{1}{\Pi_{k=1}^{r} n_k} \sigma^2_{\hat{U}}(f, D)$
Output: $\sigma^2_{\hat{U}}(f, D)$
```

Algorithm 1: Computation of $\sigma^2_{\hat{U}}(f, D)$

```
Input: data set $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$, ranking functions $f_1, f_2$;
Compute $\hat{U}(f_1, D)$ and $\hat{U}(f_2, D)$;
$\widehat{Cov}_{\hat{U}}(f_1, f_2, D) = 0$;
for $(\mathcal{Z}_1, \mathcal{Z}_2) \in \Upsilon$ do
    for $k = 1$ to $r$ do
        if $\bar{y_k} \in \mathcal{Z}_1$ then
            $V = V \cup \{((x_i, y_i), (x_i, y_i)) | y_i = \bar{y}_k\}$;
        end
        if $\bar{y_k} \in \mathcal{Z}_2$ then
            $V = V \cup \{((x_i, y_i), (x_j, y_j)) | y_i = \bar{y}_k \wedge y_j = \bar{y}_k\}$;
        end
    end
    connect $v_0 = v_s$ to all nodes in $V_1$;
    connect all nodes in $V_r$ to $v_{r+1} = v_e$;
    for $k = 1$ to $r - 1$ do
        for $v_a = ((x_1, y_1), (x_2, y_2)) \in V_k$ do
            for $v_b = ((x_3, y_3), (x_4, y_4)) \in V_{k+1}$ do
                if $f_1(x_1) < f_1(x_3) \wedge f_2(x_2) < f_2(x_4)$ then
                    $E = E \cup \{(v_a, v_b)\}$;
                end
            end
        end
    end
    $\Gamma(v_s) = 1; \forall\, v_a \in V_k, k \geq 1 : \Gamma(v_a) = 0$;
    for $k = 0$ to $r$ do
        for $v_a \in V_k$ do
            for $v_b \in V_{k+1}$ do
                if $(v_a, v_b) \in E$ then
                    $\Gamma(v_b) = \Gamma(v_b) + \Gamma(v_a)$;
                end
            end
        end
    end
    $\widehat{Cov}_{\hat{U}}(f_1, f_2, D) = \widehat{Cov}_{\hat{U}}(f_1, f_2, D) + (\Pi_{\bar{y}_l \in \mathcal{Z}_2}(n_l - 1))(\frac{\Gamma(v_e)}{\underline{n}} - \hat{U}(f_1, D)\hat{U}(f_2, D))$
end
$\widehat{Cov}_{\hat{U}}(f_1, f_2, D) = \frac{1}{\Pi_{k=1}^{r} n_k} \widehat{Cov}_{\hat{U}}(f_1, f_2, D)$
Output: $\widehat{Cov}_{\hat{U}}(f_1, f_2, D)$
```

Algorithm 2: Computation of $\widehat{Cov}_{\hat{U}}(f_1, f_2, D)$

3.2. The VUROCS package

In this section, we introduce the `R` package `VUROCS` which contains functions to calculate the VUS and its (co)variance written and maintained by the authors. The implementation is based on the algorithms proposed by [18]. The following provides a brief overview of the functionality of the `R` package.

The package `VUROCS` comes with three core functions to determine the VUS as well as variance and covariance of VUS.

(1) `VUS(y, fx)` calculates the VUS using the algorithm provided in [18] for a vector of realisations `y` and a vector of predictions `fx`.
(2) `VUSvar(y, fx)` calculates the variance of VUS using Algorithm 1 for a vector of realisations `y` and a vector of predictions `fx`.
(3) `VUScov(y, fx1, fx2)` calculates the covariance of VUS using Algorithm 2 for a vector of realisations `y`, a vector of predictions `fx1` of the first ranking function (e.g.

LGD model) and a vector of predictions `fx2` of the second ranking function under consideration.

In addition to these three functions, the package also provides implementations for the two measures CLAR and CLAR^{adj}, implemented in the functions `clar(y,fx)` and `clarAdj(y,fx)` as well as time-efficient implementations[1] of Somers' D (`SomersD(y,fx)`), Kruskal's Γ (`Kruskal_Gamma(y,fx)`), Kendall's τ_b (`Kendall_taub(y,fx)`) and Kendall's τ_c (`Kendall_tauc(y,fx)`). In addition, these functions also report asymptotic standard errors as defined in [2].

4. Application of the VUS for multi-class classification problems in LGD models

Using the functions contained in the `VUROCS` package, the aim of this section is threefold: First, Section 4.1 illustrates the drawbacks of the measures CLAR, Somers' D and Kendall's τ_b based on a hypothetical multi-class example and demonstrates the advantages of the measure VUS. Second, Section 4.2 provides basic concepts for using VUS for the purpose of LGD model validation. Finally, Section 4.3 demonstrates how the differences of discriminatory power measures may lead to different conclusions during LGD model development. The impact of using VUS instead of the discriminatory power measures described in Section 2 for modelling LGD is illustrated for a real world loss data set of a large commercial bank.

4.1. Example 1: comparison of discriminatory power measures

To demonstrate the capabilities of the discriminatory power measures presented in Section 2, Figure 2 shows the results of the following simulation: Consider an LGD model with grades $\bar{y}_1, \bar{y}_2$ and $\bar{y}_3$ with $\bar{y}_1 <_y \bar{y}_2 <_y \bar{y}_3$. For each grade, the predicted LGDs of customers belonging to this grade are drawn from normal distributions with means $\mu_{\bar{y}_1} = 1$, $\mu_{\bar{y}_2} = 2$, $\mu_{\bar{y}_3} = 3$, constant standard deviations of 0.1 and $n_1 = n_2 = n_3 = 10,000$. Starting from this ideal case, we increase $\mu_{\bar{y}_1}$ such that the perfect ordering is gradually compromised. Figure 2 then shows the values for CLAR, CLAR^{adj}, VUS, Somers' D^{rescaled}, Kendall's $\tau_b^{rescaled}$, $AUK_{tot}^{rescaled}$ and AUK (the latter on the secondary axis due to its different range and scaling), where $\mu_{\bar{y}_1}$ varies from 1 to 3.

Figure 2 demonstrates that the VUS measure is much more sensitive to distortions in the multi-class ordering than CLAR, CLAR^{adj}, Somers' D^{rescaled}, Kendall's $\tau_b^{rescaled}$, $AUK_{tot}^{rescaled}$ and AUK. As soon as the distribution of predicted LGD grades associated with observations from $\bar{y}_1$ and $\bar{y}_2$ overlap, the VUS starts to decrease sharply. Furthermore, as soon as all predicted LGD grades for observations from $\bar{y}_1$ are equal or higher than the predicted LGD grades for observations from $\bar{y}_2$, VUS becomes zero, reflecting the inability of correctly ranking observations from at least two grades. In contrast to VUS, the values obtained for CLAR, CLAR^{adj}, Somers' D, Kendall's $\tau_b^{rescaled}$, $AUK_{tot}^{rescaled}$ and AUK still indicate positive discriminatory ability in cases where the multi-class ordering is reversed. Moreover, even in cases where observations from grade $\bar{y}_1$ have highest predicted LGDs, these measures still return a value of around 0.5. While CLAR^{adj} reacts more drastic and, caused by the adjustment factor, decreases sharper than CLAR, it still fails to detect the reversal in

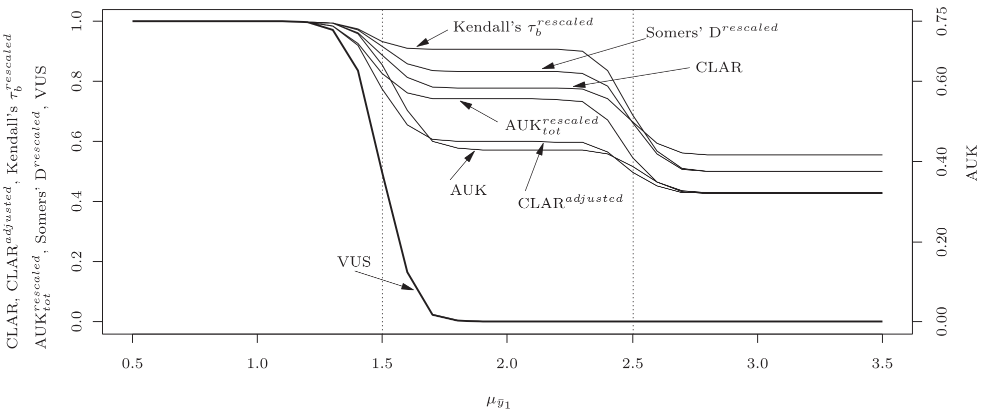

Figure 2. Discriminatory power measures for different values of $\mu_{\bar{y}_1}$.

the multi-class ordering. Similarly, AUK and $AUK_{tot}^{rescaled}$ react more intense than Kendall's $\tau_b^{rescaled}$, still they do not carry essentially different information regarding ordinal dependence by virtue of the Kendall distribution function, $K(t)$, being the defining characteristic of all AUK, $AUK_{tot}^{rescaled}$ and Kendall's $\tau_b^{rescaled}$ (see Section 2.3).

4.2. Example 2: use of VUS for LGD model validation

The CRR requires a regular cycle of model validation that includes monitoring of model performance and stability, review of model specification, and testing of model outputs against observed outcomes. One key element of LGD model validation is the assessment of discriminatory power. As only defaulted facilities for which the workout process has been finished can be considered for validation purposes, the data basis for validating LGD models is limited in many cases. For an annual model validation, the data basis should be based on all facilities for which the workout process has been finished (e.g. due to liquidation, curing etc.) in the relevant observation period. In this validation sample setup, the observation period in which the recovery process terminates remains constant, while the length of the recovery process and the year in which the default has occurred will typically be different for the individual facilities. Regarding the content of the discriminatory power analysis in the regular validation, it is of key importance from a supervisory perspective to also benchmark the current performance of the LGD model with the performance at the time of initial validation/model development based on which supervisory approval was granted. Other analyses might include the monitoring of discriminatory power over time, the testing of attaining institution-specific predefined performance thresholds or the comparison between different model versions, e.g. to analyse the gain in discriminatory power by updating the model with most recent data. To perform these discriminatory power analyses on a statistical foundation via hypothesis testing, the functions provided in the VUROCS package may be used.

For each defaulted facility in the validation sample, the estimated and realised LGDs are required to be able to perform a discriminatory power analysis. In this context, the realised LGD for each facility should be the same as the one used in calculating the long-run average LGD for the purpose of LGD estimation (Article 181(1)(a) CRR). Which estimated LGD to use is, however, less obvious since the specification of an institution's LGD model might have changed over time since a default has occurred. As the validation activities should always be related to the LGD currently in use at an institution, the estimated LGD used for validation purposes should be the LGD that would have been used for a facility to calculate own funds requirements at a time point before default if the LGD model applied at the end of the observation period had been in force at that time. The time point before default at which the predicted LGD is evaluated can be varied, e.g. between one year before default and the date of default, to monitor the pre-default stability of the LGD prediction. For facilities not existing at the chosen time point for a specific analysis, the relevant time point is the date at which the facility has been opened. Any floors on the predicted LGDs specified in Article 164(4)-(5) CRR should not be considered for the purpose of validation.

As the VUS is a U-statistic, it is asymptotically normally distributed and classical hypothesis testing theory can be applied based on its variance. In the following, we discuss several possible hypothesis tests relevant in the course of an LGD model validation that can be performed with the `VUROCS` package:

- For realisations `y` and predictions `fx`, $H_0 : \texttt{VUS}(\texttt{y}, \texttt{fx}) \geq c$, where c is a deterministic performance threshold, e.g. the VUS at the time of initial validation/model development. In this case, the null hypothesis is rejected if

$$\frac{\texttt{VUS}(\texttt{y}, \texttt{fx}) - c}{\sqrt{\texttt{VUSvar}(\texttt{y}, \texttt{fx})}} < \Phi^{-1}(\alpha),$$

  where Φ^{-1} is the quantile function of the standard normal distribution and α a predefined significance level.
- For realisations `y` and different predictions `fx1` and `fx2`, $H_0 : \texttt{VUS}(\texttt{y}, \texttt{fx2}) \geq \texttt{VUS}(\texttt{y}, \texttt{fx1})$. Here, `fx1` and `fx2` might be related to different model versions or hierarchical layers of a component-based LGD model (e.g. where the secured, unsecured and cure rate components are modelled separately). In this case, the null hypothesis is rejected if

$$\frac{\texttt{VUS}(\texttt{y}, \texttt{fx2}) - \texttt{VUS}(\texttt{y}, \texttt{fx1})}{s} < \Phi^{-1}(\alpha),$$

  where $s^2 = \texttt{VUSvar}(\texttt{y}, \texttt{fx1}) + \texttt{VUSvar}(\texttt{y}, \texttt{fx2}) - 2 \cdot \texttt{VUScov}(\texttt{y}, \texttt{fx1}, \texttt{fx2})$.
- For independent vectors of realisations `y1` and `y2` with corresponding predictions `fx1` and `fx2`, $H_0 : \texttt{VUS}(\texttt{y2}, \texttt{fx2}) \geq \texttt{VUS}(\texttt{y1}, \texttt{fx1})$. For instance, when analysing the discriminatory performance over time, $(\texttt{y1}, \texttt{fx1})$ and $(\texttt{y2}, \texttt{fx2})$ are related to different observation periods. In this case, $s^2 = \texttt{VUSvar}(\texttt{y1}, \texttt{fx1}) + \texttt{VUSvar}(\texttt{y2}, \texttt{fx2})$ in the equation above.

Note that the presented one-sided hypothesis tests are designed in a supervisor's perspective, i.e. rejection of H_0 implies the need for justification and further analysis, potentially remedial action to restore discriminatory performance.

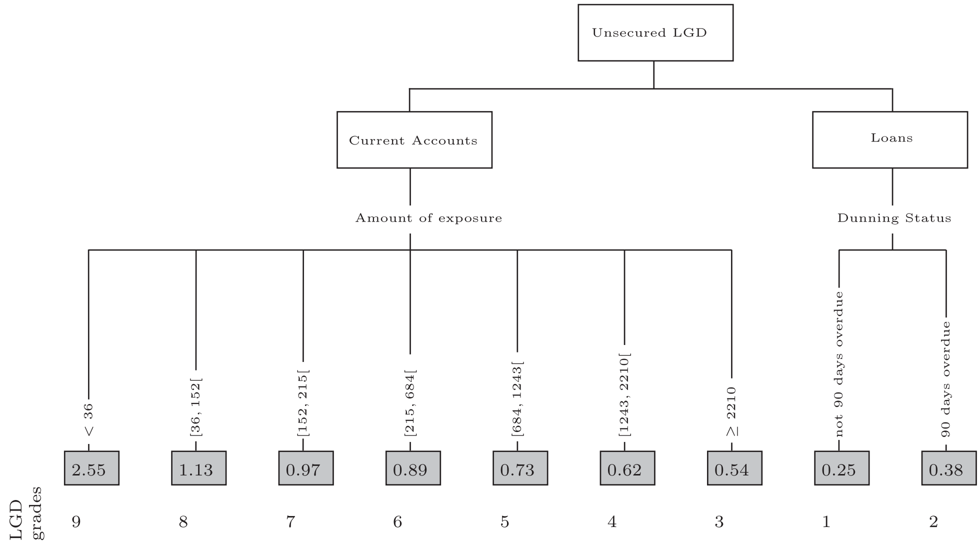

Figure 3. Overview of LGD model and the corresponding predicted LGDs for each grade

4.3. Example 3: use of VUS for LGD model development

To demonstrate the advantages of VUS for LGD model development, we present an LGD model for performing unsecured retail loans that a large commercial bank sought approval for. This LGD model was built based on loss data from an unsecured portfolio of retail facilities covering the time period 2007 to 2013. For model development, a decision tree algorithm had been used to define different segments which were intended to be homogeneous with respect to their LGD risk. An overview of the model can be found in Figure 3, where the splitting criteria as well as the predicted LGDs for each grade are displayed. The decision tree first discriminates between the product types 'current accounts' and 'loans'. The product type current accounts is further split into seven grades according to the outstanding exposure amount. Concerning the product type 'loans', the tree further differentiates between facilities with dunning status '90 days overdue' and no such dunning status. Overall, the model assigns predicted LGDs for nine different grades.

The development data sample consists of approximately 23,000 observations with a predicted and realised LGD for each observation. The top left graph of Figure 4 illustrates the density of realised LGDs. Apparently, realised LGDs are concentrated in the lower as well as in the higher LGD grades. It is a common feature (e.g. see [10,15]) that the distribution of realised LGDs has two peaks. This stems from the fact that either the customer cures and pays back the loan or the customer stays in default and does not repay the unsecured loan. Furthermore, a concentration in high loss rates is also to be expected for small ticket loans where, e.g., the recoveries may not cover the cost of the collection procedure. Figure 4 illustrates what had apparently been neglected by the bank in the model development due to the use of bivariate association measures in the judgment of discriminatory ability. The LGD model with nine grades clearly shows deficiencies as it does not adequately capture the high concentration of realised losses in low as well as in high LGD grades. Moreover, the two distributions of predicted and realised LGD do not fit well for the mid-classes.

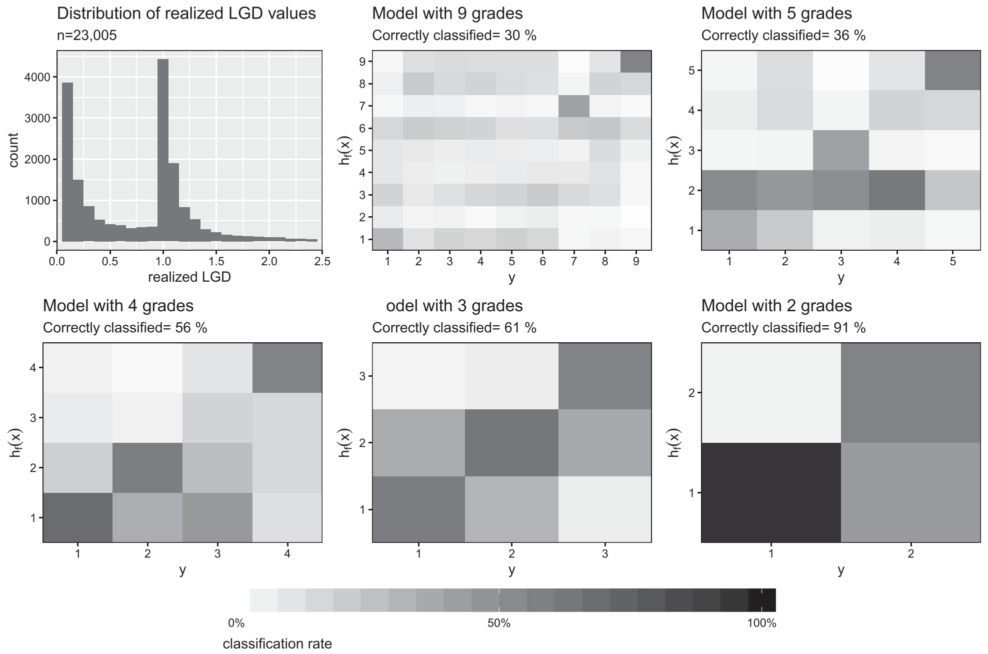

Figure 4. Distribution of realised LGDs and classification rates for LGD models with nine, five, four, three and two grades.

Table 1. VUS-optimal mapping of the original nine LGD grades to models with five, four, three and two grades.

	9 grades	5 grades	4 grades	3 grades	2 grades
LGD grades	1	1	1	1	1
	2	1	1	1	1
	3	2	1	1	1
	4	2	2	1	1
	5	2	2	1	1
	6	3	2	2	1
	7	3	2	2	1
	8	4	3	2	1
	9	5	4	3	2
VUS	0.00	0.05	0.20	0.46	0.76
Somers' $D^{rescaled}$	0.71	0.70	0.69	0.70	0.76
$CLAR^{adj}$	0.58	0.66	0.54	0.60	0.55

This is reflected in a value of VUS of approximately zero as reported in Table 1 and could have therefore been detected without analysing the realised loss distributions by using a discriminatory power measure suitable for the multi-class setting.

Consider now as a toy example the most simple way of following a VUS-led completion of the model development at this stage. To address the problem of high discrepancies of predicted and realised LGDs in mid-classes, the model might be improved by merging grades. Table 1 presents the results of optimal models with nine, five, four, three and two grades, where the model selection is based on the maximisation of VUS. Table 1 also displays the levels of the discriminatory power measures for the full model with nine grades as well as for the reduced models. Obviously, the discriminatory power is expected to increase when the number of grades is reduced. While the increase in discriminatory ability when merging LGD grades can be observed in the case of VUS, Somers' D^{rescaled} and CLAR^{adj} seem to be more or less indifferent between the full and the reduced models. Thus, we observe that both Somers' D^{rescaled} and CLAR^{adj} do not give clear indication regarding the optimal modelling choice.

The increase in discriminatory ability when merging LGD grades can also be observed in Figure 4, which compares the classification rates in the LGD grades across the different models. Based on the VUS, a model with three to four grades seems to be preferable as one can observe a sharp increase in VUS while retaining a few grades for discrimination even in a complex to model unsecured portfolio. As alternative decision criteria to the VUS itself, VUS-based measures that consider the dimensionality r and the value of $1/r!$ for VUS in case of a random classifier for r LGD grades give additional insight. While $(\hat{U}(f,D))^{1/r}$ can be interpreted as a geometric mean of the true class fractions, a VUS-based accuracy ratio is $(\hat{U}(f,D) - 1/r!)/(1 - 1/r!)$ and taking its rth root makes the volume comparable among dimensions. In the present example, a model with three grades is favoured by the latter criterion.

The results also give clear indications on the suitability of the explanatory variables in the full model. In particular, they highlight that the facility-level days-past-due status is a poor explanatory variable for LGD and that, in accordance with the practical considerations, the exposure amount does have significant discriminatory ability. Still, for an LGD model actually seeking regulatory approval, further explanatory variables would need to be checked for their contribution to risk differentiation such as bank-internal behavioural factors (e.g. previous repayments, remaining tenor, forborne principal payments, previous defaults, internal or external collection), contract characteristics and general obligor characteristics (e.g. months on book, personal information) and external categories (e.g. macroeconomic information).

5. Conclusion and discussion

Based on applications from the area of banking supervision, this paper establishes that bivariate (ordinal) measures of association are inadequately able to assess the discriminatory power of multi-class LGD models. The fact that these measures solely consider bivariate orderings is a major restriction which impairs their reliability for the purpose of LGD model validation and development. As an alternative to the widely used measures CLAR, Somers' D and Kendall's τ_b or the more recent proposals for measures such as AUK or the related novel ordinal measure of association AUK_{tot}, we suggest using the VUS for measuring discriminatory power of LGD models. VUS appropriately reflects the multi-class ordering, in this way allowing credible judgment about the discriminatory ability based on classical statistical methodology via its relation to U-statistics. The R-package VUROCS provides an efficient implementation to calculate the VUS as well as the (co)variance of VUS for multi-category classification problems that can be directly applied by practitioners in LGD model validation and development.

More involved employment of the measure VUS for the purpose of LGD model development as a model selection criterion is the subject of future scientific work. Instead of acting as an ex-post criterion to compare LGD models, an extension would be to directly use the VUS as the optimisation criterion within the algorithms to select explanatory variables and to identify the optimum number of LGD grades. Also in the real-world example provided in Section 4.3, this would open the possibility to elicit models with even higher heterogeneity between grades according to VUS, allowing an even more granular and credible risk differentiation.

Note

1. Simulations with $k = 2, 6, 10, \ldots, 58$ classes and $n = 100, 123, 146, \ldots, 1000$ observations indicate that the implementations of Somers' D result in an average decrease in computation time of around 96% (measured using `R`'s `microbenchmark` package with 1000 repetitions) compared to the implementation found in `R`'s `DescTools` package (Version 0.99.24).

Acknowledgments

The authors are grateful to Manveer Mangat who supported coding and drafting in early stages of the project. The paper expresses the authors' views and does not represent official positions of the Oesterreichische Nationalbank.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Basel Committee on Banking Supervision, *International convergence of capital measurements and capital standards: A revised framework*, Basel Committee on Banking Supervision, 2006.

[2] M.B. Brown and J.K. Benedetti, *Sampling behavior of tests for correlation in two-way contingency tables*, J. Am. Stat. Assoc. 72 (1977), pp. 309–315.

[3] S. Dreiseitl, L. Ohno-Machado, and M. Binder, *Comparing three-class diagnostic tests by three-way ROC analysis*, Med. Decis. Making 20 (2000), pp. 323–331.

[4] B. Engelmann, E. Hayden, and D. Tasche, *Measuring the discriminative power of rating systems*, Discussion paper, Series 2: Banking and financial supervision, Deutsche Bundesbank, 2003.

[5] European Banking Authority, *EBA results from the credit risk benchmarking exercise 2019 (HDP+LDP)*, (2020). Available at https://eba.europa.eu/sites/default/documents/files/document_library/Publications/Reports/2020/EBA%20Report%20-%20Results%20from%20the%202019%20Credit%20Risk%20Benchmarking%20Exercise.pdf.

[6] European Central Bank, *ECB guide to internal models*, (2019). Available at https://www.bankingsupervision.europa.eu/ecb/pub/pdf/ssm.guidetointernalmodels_consolidated_20191097fd49fb08.en.pdf.

[7] European Central Bank, *Instructions for reporting the validation results of internal models. IRB Pillar I model for credit risk*, (2019). Available at https://www.bankingsupervision.europa.eu/banking/tasks/internal_models/shared/pdf/instructions_validation_reporting_credit_risk.en.pdf.

[8] C. Genest and J.-C. Boies, *Detecting dependence with Kendall plots*, Amer. Statist. 57 (2003), pp. 275–284.

[9] D.J. Hand and R.J. Till, *A simple generalization of the area under the ROC curve for multiple class classification problems*, Mach. Learn. 45 (2001), pp. 171–186.

[10] D. Li, R. Bhariok, S. Keenan, and S. Santilli, *Validation techniques and performance metrics for loss given default models*, The Journal of Risk Model Validation 3 (2009).

[11] D. Mossman, *Three-way ROCs*, Med. Decis. Making 19 (1999), pp. 78–89.

[12] C.T. Nakas and C.T. Yiannoutsos, *Ordered multiple-class ROC analysis with continuous measurements*, Stat. Med. 23 (2004), pp. 3437–3449.

[13] R.B. Nelsen, *An Introduction to Copulas*, New York: Springer, 2006.

[14] Oesterreichische Nationalbank, *Guidlines on credit risk management. Rating models and validation*, Oesterreichische Nationalbank and Austrian Financial Market Authority, 2004.

[15] B. Ozdemir and P. Miu, *Basel II Implementation. A Guide to Developing and Validating a Compliant Internal Risk Rating System*, New York: McGraw-Hill, 2009.

[16] S. Scandizzo, *Loss given default models*, in *The Validation of Risk Models: A Handbook for Practitioners*, Applied Quantitative Finance, Palgrave Macmillan, 2016, pp. 78–91.

[17] D. Tasche, *Validation of internal rating systems and PD estimates*, in *The Analytics of Risk Model Validation*, Quantitative Finance Series, Academic Press, 2008, pp. 169–196.

[18] W. Waegeman, B. De Baets, and L. Boullart, *On the scalability of ordered multi-class ROC analysis*, Comput. Stat. Data Anal. 52 (2008), pp. 3371–3388.

[19] A. Vexler, X. Chen, and A.D. Hutson, *Dependence and independence: Structure and inference*, Stat. Methods Med. Res. 26 (2017), pp. 2114–2132.

[20] A. Vexler, G. Afendras, and M. Markatou, *Multi-panel Kendall plot in light of an ROC curve analysis applied to measuring dependence*, Statistics 53 (2019), pp. 417–439.