Nonparametric predictive inference for the validation of credit rating systems

T. Coolen-Maturi†

Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK.

E-mail: tahani.maturi@durham.ac.uk

F.P.A. Coolen

Department of Mathematical Sciences, Durham University, Durham, DH1 3LE, UK.

E-mail: frank.coolen@durham.ac.uk

Summary. Credit rating or credit scoring systems are important tools for estimating the obligor's credit worthiness and for providing an indication of the obligor's future status. The discriminatory power of a credit rating or credit scoring system refers to its ex ante ability to distinguish between two or more classes of borrowers. One of the most popular tools for the validation of the power of credit rating or credit scoring models to distinguish between two (or more) classes of borrowers is the receiver operating characteristic ROC curve (hypersurface) and its widely used overall summary, the area (hypervolume) under the curve (hypersurface). As the end goal of building such models is to predict and quantify uncertainty about future loans, prediction methods are especially valuable in this context. To this end, nonparametric predictive inference (NPI) is a promising candidate for such inference as it is a frequentist statistical method that is explicitly aimed at using few modelling assumptions, enabled through the use of lower and upper probabilities to quantify uncertainty. The aim of this paper is to introduce NPI for ROC analysis within a banking context, to which end novel results on ROC hypersurfaces for more than three groups are presented. Examples are provided to illustrate the method.

Keywords: Nonparametric predictive inference, credit rating systems, receiver operating characteristic curve, hypervolume under the ROC hypersurface.

1. Introduction

With the financial crisis, many banks and building societies are facing challenges over loan repayment, which makes decisions on approval of loans crucial. To this end, many statistical methods have been introduced to classify an applicant into different classes based on previous loan customers' records (Hand and Henley, 1997). The problem of assessing the accuracy of these classifiers is important and thus several statistical tools have been developed, such as ROC (receiver operating characteristic) curve and Gini coefficient (Crook et al., 2007). In particular, the ROC approach is a well-known tool for assessing the discriminating ability of credit rating systems (Crook et al., 2007; Xanthopoulos and Nakas, 2007).

†Corresponding author.

As the reality is that the banks and building societies are interested in evaluating future loan customers, it will be of interest to consider statistical methods for prediction rather than estimation. Classical methods often focus on estimation to make inferences about future loans, and they often require the underlying distributions of borrowers to be known, which is unrealistic in practice. To this end, Nonparametric Predictive Inference (NPI) is a promising candidate for such inference as it does not require any assumed distributions and the inference itself is explicitly predictive.

NPI is a statistical method based on Hill's assumption $A_{(n)}$ (Hill, 1968), which gives a direct conditional probability for a future observable random quantity, conditional on observed values of related random quantities (Augustin and Coolen, 2004). $A_{(n)}$ does not assume anything else, and can be interpreted as a post-data assumption related to exchangeability (De Finetti, 1974). Inferences based on $A_{(n)}$ are predictive and nonparametric, and can be considered suitable if there is hardly any knowledge about the random quantity of interest, other than the n observations, or if one does not want to use such information, e.g. to study effects of additional assumptions underlying other statistical methods. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but it provides bounds for probabilities via the 'fundamental theorem of probability' (De Finetti, 1974). These bounds are lower and upper probabilities in imprecise probability theory (Augustin and Coolen, 2004). An informal interpretation for lower and upper probabilities, is that a lower probability reflects the evidence in favour of the event of interest while an upper probability reflects the evidence against the event of interest.

NPI has been introduced for several application areas including reliability, survival analysis, operations research and finance (Coolen, 2011; Baker et al., 2017; He et al., 2018). A short introduction to NPI and its applications is given by Coolen (2011). NPI has also been introduced for assessing the accuracy of a classifier's ability to discriminate between two outcomes (or two groups) for binary data (Coolen-Maturi et al., 2012a) and for diagnostic tests with ordinal observations (Elkhafifi and Coolen, 2012) and with real-valued observations (Coolen-Maturi et al., 2012b). Recently, Coolen-Maturi et al. (2014) generalized the results by Coolen-Maturi et al. (2012b) by introducing NPI for three-group ROC analysis, with real-valued observations, to assess the ability of a diagnostic test to discriminate among three ordered classes or groups. Coolen-Maturi (2017b) generalized the results by Elkhafifi and Coolen (2012) by proposing NPI for three-group ROC analysis with ordinal outcomes. This paper generalizes the methods by Coolen-Maturi (2017b) for more than three ordered classes or groups. The aim of this paper is to introduce NPI for ROC analysis within a banking context. In particular to assess the discriminatory power of a credit rating or credit scoring system refering to its ex ante ability to distinguish between multiple ordered classes of borrowers.

This paper is organised as follows. Section 2 introduces NPI for ROC analysis for ordinal outcomes within a banking context including novel results on ROC hypersurfaces for more than three groups. Two examples are provided in Section 4 and some concluding remarks are given in Section 5.

Table 1. Ordinal test data

	Credit scoring model outcomes notation								
Status	C_1	...	C_{k_1}	...	C_{k_2}	...	$C_{k_{G-1}}$	...	C_K	Total
Y^1	n_1^1	...	$n_{k_1}^1$	...	$n_{k_2}^1$	...	$n_{k_{G-1}}^1$	...	n_K^1	n^1
Y^2	n_1^2	...	$n_{k_1}^2$	...	$n_{k_2}^2$	...	$n_{k_{G-1}}^2$	...	n_K^2	n^2
$\vdots$	$\vdots$		$\vdots$		$\vdots$		$\vdots$		$\vdots$	$\vdots$
Y^G	n_1^G	...	$n_{k_1}^G$	...	$n_{k_2}^G$	...	$n_{k_{G-1}}^G$	...	n_K^G	n^G
Total	n_1	...	n_{k_1}	...	n_{k_2}	...	$n_{k_{G-1}}$	...	n_K	n

2. NPI for ROC analysis for ordinal outcomes

We consider a credit scoring model with ordinal outcomes, where the outcome for each borrower indicates one of $K \geq 3$ ordered classes, denoted by C_1 to C_K and representing a decreasing level of credit worthiness (e.g. from excellent to poor credit worthiness). We assume that the data available are on borrowers in G ordered groups according to known status indicated by $Y^1, Y^2, \ldots, Y^G$. We assume that there are $G-1$ cut-off points (or thresholds) $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, ..., K\}$ such that a score in classes $\{C_1, \ldots, C_{k_1}\}$ is interpreted as indication that a borrower is belonging to the first group (status), a score in classes $\{C_{k_{i-1}+1}, \ldots, C_{k_i}\}$ as indication that a borrower is belonging to the ith group, and finally a score in classes $\{C_{k_{K-1}+1}, \ldots, C_K\}$ as indication that a borrower is belonging to the final group (the Gth status). The notation for the numbers of borrowers for each combination of status and model outcomes is given in Table 1, where n is the total number of borrowers, n_j^i is the number of borrowers from group i in class C_j and n^i is total number of borrowers from group i.

For thresholds $k_1 < k_2 < \ldots < k_{G-1}$, the probability of correct classification of a borrower from group Y^i is $p_i(k_{i-1}, k_i) = P(Y^i \in \{C_{k_{i-1}+1}, \ldots, C_{k_i}\})$, $i = 1, 2, \ldots, G$. The ROC hypersurface can be constructed by plotting these probabilities of correct classification $\{p_i(k_{i-1}, k_i), i = 1, 2, \ldots, G\}$ for all $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, \ldots, K\}$. The empirical estimators of these probabilities are $\hat{p}_i(k_{i-1}, k_i) = (1/n^i) \sum_{j=k_{i-1}+1}^{k_i} n_j^i$, for $i = 1, 2, \ldots, G$, where for simplicity of notation we assume $k_0 = 0$ and $k_G = K$, which in turn form the empirical ROC hypersurface, denoted by $\widehat{\text{ROCs}}$.

The hypervolumes under the ROC hypersurface (VUHS) can be used as a global measure of the discriminatory ability of the test under consideration. It can take values from 0 to 1, where a value of about $1/G!$ would occur if the observations from the G groups would fully overlap, in such a way that the credit scoring model would perform no better than a random allocation of subjects to the G groups. If there is a perfect separation of the test results for the G groups, then $VUHS = 1$. Nakas and Yiannoutsos (2004) presented the hypervolume under the ROC hypersurface for real-valued data, in this paper we utilize their findings using a latent variable representations of ordinal data to obtain the empirical estimator of the hypervolume under the ROC hypersurface. For the special case when $G = 4$, the empirical estimator of the hypervolume under the ROC hypersurface, $\widehat{VUHS}$, is given in the appendix.

For NPI, we are interested in the inference about the next obligor's status, where we consider one next obligor from each group, that is $Y_{n^1+1}^1, Y_{n^2+1}^2, \ldots, Y_{n^G+1}^G$. As the $A_{(n)}$

assumption is only suitable for real-valued data, a latent variable representation has been used for inference about an ordinal random quantity, similar to the method presented by Coolen-Maturi (2017b) for the case $G = 3$. Thus in this section we generalise the results in Coolen-Maturi (2017b) for more than three groups, using the idea of the latent variables representations. The NPI lower and upper probabilities of correct classification, for the thresholds $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, 2, \ldots, K\}$, and for $i = 1, 2, \ldots, G$, where $k_0 = 0$ and $k_G = K$, are

$$\begin{aligned}\underline{p}_i(k_{i-1}, k_i) &= \underline{P}(Y^i_{n^i+1} \in \{C_{k_{i-1}+1}, \ldots, C_{k_i}\}) \\ &= \begin{cases} \dfrac{1}{n^i+1} \displaystyle\sum_{j=k_{i-1}+1}^{k_i} n^i_j & \text{if } i = 1 \text{ or } i = G \\ \dfrac{1}{n^i+1} \left(-1 + \displaystyle\sum_{j=k_{i-1}+1}^{k_i} n^i_j \right)^+ & \text{if } i = 2, 3 \ldots, G-1 \end{cases} \end{aligned} \tag{1}$$

$$\overline{p}_i(k_{i-1}, k_i) = \overline{P}(Y^i_{n^i+1} \in \{C_{k_{i-1}+1}, \ldots, C_{k_i}\}) = \frac{1}{n^i+1} \left(1 + \sum_{j=k_{i-1}+1}^{k_i} n^i_j \right) \tag{2}$$

where $(x)^+ = \max(x, 0)$.

We can define the following ROC hypersurfaces with VUHS values equal to the infimum and supremum of the VUHS values for all NPI-based ROC hypersurfaces. The equality of the VUHS and the probability of correctly ordered observations enables us to define lower and upper ROC hypersurfaces in line with the optimization procedures, similar to the one described by Coolen-Maturi (2017b), to obtain $\underline{n}^i_j$ and $\overline{n}^i_j$, $i = 2, \ldots, G-1$. For simplicity of notation, for the first and last groups, let $\underline{n}^1_j = \overline{n}^1_j = n^1_j$ and $\underline{n}^G_j = \overline{n}^G_j = n^G_j$. These lower and upper ROC hypersurface are defined as follows.

The NPI lower ROC hypersurface, $\underline{ROCs}$, goes through the points $\big\{\big(p_1(k_0, k_1),\ p^*_i(k_{i-1}, k_i),\ p_G(k_{G-1}, k_G)\big) : p_1(k_0, k_1) \in \big[\overline{p}_1(k_0, k_1) - \overline{p}_1(k_0, k_1 - 1)\big], p_G(k_{G-1}, k_G) \in \big[\underline{p}_G(k_{G-1}, k_G) - \underline{p}_G(k_{G-1} + 1, k_G)\big],\ k_1 < k_2 < \ldots < k_{G-1} \in \{1, \ldots, K\}\big\}$, where $p^*_i(k_{i-1}, k_i) = (n^i + 1)^{-1} \sum_{j=k_{i-1}+1}^{k_i} \underline{n}^i_j$, $i = 2, \ldots, G-1$. The NPI upper ROC hypersurface, $\overline{ROCs}$, goes through the points $\big\{\big(p_1(k_0, k_1), p^{**}_i(k_{i-1}, k_i), p_G(k_{G-1}, k_G)\big) : p_1(k_0, k_1) \in \big[\underline{p}_1(k_0, k_1) - \underline{p}_1(k_0, k_1 - 1)\big], p_G(k_{G-1}, k_G) \in \big[\overline{p}_G(k_{G-1} - 1, k_G) - \overline{p}_G(k_{G-1}, k_G)\big],\ k_1 < k_2 < \ldots < k_{G-1} \in \{1, \ldots, K\}\big\}$, where $p^{**}_i(k_{i-1}, k_i) = (n^i + 1)^{-1} \sum_{j=k_{i-1}}^{k_i} \overline{n}^i_j$, $i = 2, \ldots, G-1$.

To present the hypervolumes under the lower and upper ROC hypersurfaces, we need to introduce further notation. Let $S^d = \{X \subset S : |X| = d\}$ denote the set of all subsets of $S = \{1, 2, \ldots, G\}$ of size d, where $d = 0, 1, \ldots, G$. That is we have $\binom{G}{d}$ subsets of S of size d. Note that the empty set corresponds to $d = 0$. Similarly we can define $S^{d_1}_1 = \{X \subset S : \{1, G\} \subset X \wedge |X| = d_1\}$ where $d_1 = 2, 3, \ldots, G$, and $S^{d_2}_2 = \{X \subset S : \{2, \ldots, G-1\} \subset X \wedge |X| = d_2\}$, where $d_2 = G-2, G-1, G$.

The hypervolumes under the NPI lower and upper ROC hypersurface, which are equal to the NPI lower and upper probabilities for the event $(Y^1_{n^1+1} < Y^2_{n^2+1} < \ldots < Y^G_{n^G+1})$,

respectively, are

$$\underline{VUHS} = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{i_1=1}^{K-G+1} \sum_{i_2=i_1+1}^{K-G+2} \cdots \sum_{i_{G-1}=i_{G-2}+1}^{K-1} \sum_{i_G=i_{G-1}+1}^{K} \prod_{g=1}^{G} \underline{n}_{i_g}^{g} \tag{3}$$

$$\overline{VUHS} = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{d_2=G-2}^{G} \left[\sum_{J \in S_2^{d_2}} \sum_{i_{J[1]}=1}^{K} \sum_{i_{J[2]}=i_{J[1]}}^{K} \cdots \sum_{i_{J[d_2]}=i_{J[d_2-1]}}^{K} \prod_{g \in J} \overline{n}_{i_g}^{g} \right] \tag{4}$$

where $\sum_{J \in S_2^{d_2}}$ denote the sum over all the subsets in $S_2^{d_2}$, and $J[j]$ refers to the jth element in J, $j = 1, 2, \ldots, d_2$.

2.1. Lower and upper envelopes of the set of NPI-based ROC hypersurfaces

One may want to avoid the numerical optimisations (especially for a large data set with a large number of categories) required to derive the NPI lower and upper ROC hypersurfaces above, by using envelopes as approximations, benefitting from the fact that they are available in simple analytical expressions as given in below. These envelopes provide lower and upper bounds for the NPI lower and upper ROC surfaces, which provide some further information about the quality of the approximations.

It is easy to show that the lower bound for the NPI lower ROC hypersurface, $\underline{ROCs}^L$, goes through the points $\big\{\big(p_1(k_0,k_1), \underline{p}_i(k_{i-1},k_i), p_G(k_{G-1},k_G)\big) : p_1(k_0,k_1) \in \big[\overline{p}_1(k_0,k_1) - \overline{p}_1(k_0,k_1-1)\big], p_G(k_{G-1},k_G) \in \big[\underline{p}_G(k_{G-1},k_G) - \underline{p}_G(k_{G-1}+1,k_G)\big], k_1 < k_2 < \ldots < k_{G-1} \in \{1,\ldots,K\}\big\}$, where $\underline{p}_i(k_{i-1},k_i)$ is obtained from (1). On the other hand, the upper bound for the NPI upper ROC hypersurface, $\overline{ROCs}^U$, goes through the points $\big\{\big(p_1(k_0,k_1), \overline{p}_i(k_{i-1}-1,k_i), p_G(k_{G-1},k_G)\big) : p_1(k_0,k_1) \in \big[\underline{p}_1(k_0,k_1) - \underline{p}_1(k_0,k_1-1)\big], p_G(k_{G-1},k_G) \in \big[\overline{p}_G(k_{G-1}-1,k_G) - \overline{p}_G(k_{G-1},k_G)\big], k_1 < k_2 < \ldots < k_{G-1} \in \{1,\ldots,K\}\big\}$, where $\overline{p}_i(k_{i-1}-1,k_i)$ is obtained from (2).

Obtaining the NPI lower and upper probabilities for specific orderings of future observations and thus the corresponding hypervolume under the ROC hypersurface is computationally intensive, while having bounds for both the lower and upper hypervolumes is likely to be sufficient. Below we present such bounds, the fact that these lower and upper bounds have explicit formulas makes further optimisation unnecessary. In this section, we present the main results without formal proofs, as these follow similar steps as presented for the three-group case by Coolen-Maturi (2017b).

The NPI lower bound for the lower hypervolume under the ROC hypersurface is

$$\underline{VUHS}^L = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{d_1=2}^{G} \left[(-1)^{G-d_1} \sum_{J \in S_1^{d_1}} \sum_{i_{J[1]}=1}^{K-G+J[1]} \sum_{i_{J[2]}=i_{J[1]}+J[2]-J[1]}^{K-G+J[2]} \cdots \right.$$
$$\left. \cdots \sum_{i_{J[d_1]}=i_{J[d_1-1]}+J[d_1]-J[d_1-1]}^{K-G+J[d_1]} \prod_{g \in J} n_{i_g}^{g} \right] \tag{5}$$

where $\sum_{J\in S_1^{d_1}}$ denote the sum over all the subsets in $S_1^{d_1}$, and $J[j]$ refers to the jth element in J, $j = 1, 2, \ldots, d_1$. The NPI upper bound for the upper hypervolume under the ROC hypersurface is

$$\overline{VUHS}^U = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{d=0}^{G} \left[\sum_{J\in S^d} \sum_{i_{J[1]}=1}^{K} \sum_{i_{J[2]}=i_{J[1]}}^{K} \cdots \sum_{i_{J[d]}=i_{J[d-1]}}^{K} \prod_{g\in J} n_{i_g}^g \right] \tag{6}$$

where $\sum_{J\in S^d}$ denote the sum over all the subsets in S^d, and $J[j]$ refers to the jth element in J, $j = 1, 2, \ldots, d$.

Similarly, it is easy to show that the upper bound for the NPI lower ROC hypersurface, $\underline{ROCs}^U$, goes through the points $\big\{\big(p_1(k_0,k_1), \tilde{p}_i(k_{i-1}+1,k_i), p_G(k_{G-1},k_G)\big) : p_1(k_0,k_1) \in \big[\overline{p}_1(k_0,k_1) - \overline{p}_1(k_0,k_1-1)\big], p_G(k_{G-1},k_G) \in \big[\underline{p}_G(k_{G-1},k_G) - \underline{p}_G(k_{G-1}+1,k_G)\big], k_1 < k_2 < \ldots < k_{G-1} \in \{1,\ldots,K\}\big\}$, where $\tilde{p}_i(k_{i-1}+1,k_i) = (n^i+1)^{-1}\sum_{j=k_{i-1}+1}^{k_i} n_j^i$. On the other hand, the lower bound for the NPI upper ROC hypersurface, $\overline{ROCs}^L$, goes through the points $\big\{\big(p_1(k_0,k_1), \tilde{p}_i(k_{i-1},k_i), p_G(k_{G-1},k_G)\big) : p_1(k_0,k_1) \in \big[\underline{p}_1(k_0,k_1) - \underline{p}_1(k_0,k_1-1)\big], p_G(k_{G-1},k_G) \in \big[\overline{p}_G(k_{G-1}-1,k_G) - \overline{p}_G(k_{G-1},k_G)\big], k_1 < k_2 < \ldots < k_{G-1} \in \{1,\ldots,K\}\big\}$, where $\tilde{p}_i(k_{i-1},k_i) = (n^i+1)^{-1}\sum_{j=k_{i-1}}^{k_i} n_j^i$.

The NPI upper bound for the lower hypervolume under the ROC hypersurface is

$$\underline{VUHS}^U = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{i_1=1}^{K-G+1} \sum_{i_2=i_1+1}^{K-G+2} \cdots \sum_{i_{G-1}=i_{G-2}+1}^{K-1} \sum_{i_G=i_{G-1}+1}^{K} \prod_{g=1}^{G} n_{i_g}^g \tag{7}$$

The NPI lower bound for the upper hypervolume under the ROC hypersurface is

$$\overline{VUHS}^L = \frac{1}{\prod_{g=1}^{G}(n^g+1)} \sum_{d_2=G-2}^{G} \left[\sum_{J\in S_2^{d_2}} \sum_{i_{J[1]}=1}^{K} \sum_{i_{J[2]}=i_{J[1]}}^{K} \cdots \sum_{i_{J[d_2]}=i_{J[d_2-1]}}^{K} \prod_{g\in J} n_{i_g}^g \right] \tag{8}$$

where $\sum_{J\in S_2^{d_2}}$ denote the sum over all the subsets in $S_2^{d_2}$, and $J[j]$ refers to the jth element in J, $j = 1, 2, \ldots, d_2$.

For the sake of illustration, we have provided in the appendix the four formulas, from equations (5), (6), (7) and (8), for the special case when $G = 4$, along with the empirical estimator.

Finally, it is also worth mentioning here the three special cases: (1) when we have two groups and only two categories, (2) two groups with any number of categories and (3) three groups with any number of categories. In these cases the results are identical to those obtained by Coolen-Maturi et al. (2012a), Elkhafifi and Coolen (2012) and Coolen-Maturi (2017b), respectively. In Section 4 we use the conventional notation AUC for the area under the ROC curve (two-group case), and VUS for the volume under the ROC surface (three-group case). For example, $\underline{VUS}$ and $\overline{VUS}$ refer to the NPI lower and upper volumes under the lower and upper ROC surfaces, and the corresponding lower and upper envelopes as $(\underline{VUS}^L, \underline{VUS}^U)$ and $(\overline{VUS}^L, \overline{VUS}^U)$, respectively.

2.2. The NPI-based optimal decision thresholds

The selection of the optimal cut-off points $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, ..., K\}$, is an important aspect of defining the credit scoring model and analysing its quality. One approach is maximization of Youden's index (Youden, 1950), which for a continuous diagnostic test was introduced by Nakas et al. (2010). Similarly, we can define Youden's index for an ordinal G-group classifier as

$$J(k_1, k_2, \ldots, k_{G-1}) = \sum_{i=1}^{G} p_i(k_{i-1}, k_i) \tag{9}$$

Using this index, the optimal cut-off points are the values of $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, ..., K\}$ which maximise $J(k_1, k_2, \ldots, k_{G-1})$. This index $J(k_1, k_2, \ldots, k_{G-1})$ is equal to 1 if the G groups fully overlap, while $J(k_1, k_2, \ldots, k_{G-1}) = G$ if the G groups are perfectly separated. The empirical estimator for $J(k_1, k_2, \ldots, k_{G-1})$ is obtained by replacing these probabilities by their corresponding empirical estimators,

$$\widehat{J}(k_1, k_2, \ldots, k_{G-1}) = \sum_{i=1}^{G} \hat{p}_i(k_{i-1}, k_i) \tag{10}$$

The NPI lower and upper probabilities of correct classification for all G groups, Equations (1) and (2), can be used to obtain the NPI lower and upper bounds for Youden's index as follows,

$$\underline{J}(k_1, k_2, \ldots, k_{G-1}) = \sum_{i=1}^{G} \underline{p}_i(k_{i-1}, k_i) \tag{11}$$

$$\overline{J}(k_1, k_2, \ldots, k_{G-1}) = \sum_{i=1}^{G} \overline{p}_i(k_{i-1}, k_i) \tag{12}$$

These generalize the three-group Youden's index presented by Coolen-Maturi (2017b). Note that there is a constant difference between the NPI lower and upper Youden's indices which implies that both will be maximised at the same values of $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, ..., K\}$. It is further easy to show that, for all $k_1 < k_2 < \ldots < k_{G-1}$, $\underline{J}(k_1, k_2, \ldots, k_{G-1}) \leq \widehat{J}(k_1, k_2, \ldots, k_{G-1}) \leq \overline{J}(k_1, k_2, \ldots, k_{G-1})$, where $\widehat{J}(k_1, k_2, \ldots, k_{G-1})$ is the empirical estimate of Youden's index. These inequalities do not imply that the empirical estimate of Youden's index is maximal for the same values of $k_1 < k_2 < \ldots < k_{G-1}$ in $\{1, ..., K\}$ as the NPI lower and upper Youden's indices, but we expect that in many situations the maxima will be attained as the same values, in particular for small K.

3. Simulation

In this section, a simulation study is provided to evaluate and compare the performance of the proposed NPI method with the classical empirical method. We consider five cases (denoted by Case A to Case E) from two different distributions, namely the Uniform and the Logit-Normal distribution. These five cases are constructed to represent different (decreasing) levels of overlapping between the three ordered groups ($G = 3$), where the three groups in Case A are drawn from relatively separated distrubtions until Case E where the three groups are drawn from the same distribution (totally overlap). For each case and per distribution, we consider two data scenarios ($n^1 = n^2 = n^3 = 100$ and $n^1 = 180, n^2 = 100, n^3 = 10$) for 10 and 20 classes ($K = 10, 20$). The later data scenario ($n^1 = 180, n^2 = 100, n^3 = 10$) is of a particular interest, because in practice it is quite likely that numbers in the different groups differ substantially. For example, individuals in group 3 may lead to severe problems, but one would probably not have many of such individuals in the data. We use the cut-points 0.1(0.1)0.9 to categorise the simulated data into $K = 10$ categories, and the cut-points 0.05(0.05)0.95 to categorise the simulated values into $K = 20$ categories or classes. For the Logit-Normal distribution, the R package `logitnorm` has been used (Wutzler, 2018). All results in this section are based on $N = 10,000$ simulations.

For the Uniform distribution, the five cases are defined as follows, also displayed in Fig. 1,
Case A: $Y^1 \sim U(0, 1/3)$, $Y^2 \sim U(1/3, 2/3)$, $Y^3 \sim U(2/3, 1)$
Case B: $Y^1 \sim U(0, 1/3)$, $Y^2 \sim U(0.3, 0.65)$, $Y^3 \sim U(0.6, 1)$
Case C: $Y^1 \sim U(0, 1/3)$, $Y^2 \sim U(1/4, 2/3)$, $Y^3 \sim U(1/2, 1)$
Case D: $Y^1 \sim U(0, 0.4)$, $Y^2 \sim U(1/4, 0.7)$, $Y^3 \sim U(0.4, 1)$
Case E: $Y^1 \sim U(0, 1)$, $Y^2 \sim U(0, 1)$, $Y^3 \sim U(0, 1)$

and for the Logit-Normal distribution, the five cases are defined as follows, also displayed in Fig. 2,
Case A: $Y^1 \sim LN(-2, 0.4)$, $Y^2 \sim LN(0, 0.3)$, $Y^3 \sim LN(2, 0.4)$
Case B: $Y^1 \sim LN(-1.5, 0.4)$, $Y^2 \sim LN(0, 0.3)$, $Y^3 \sim LN(1.5, 0.4)$
Case C: $Y^1 \sim LN(-1, 0.4)$, $Y^2 \sim LN(0, 0.3)$, $Y^3 \sim LN(1, 0.4)$
Case D: $Y^1 \sim LN(-1.2, 1)$, $Y^2 \sim LN(0, 0.5)$, $Y^3 \sim LN(1.2, 1)$
Case E: $Y^1 \sim LN(0, 1)$, $Y^2 \sim LN(0, 1)$, $Y^3 \sim LN(0, 1)$

For each case we simulate $n^1 + 1, n^2 + 1, n^3 + 1$ data observations, where n^1, n^2, n^3 observations will be used to find the optimal thresholds k_1 and k_2 that maximise the empirical, the lower and the upper Youden's index (equations (10), (11) and (12), respectively). Then the future observations (one per group) will be used to evaluate the proposed and the empirical methods, that whether these future observations are correctly classified. That is whether $Y^1_{n^1+1} \in \{C_1, \ldots, C_{k_1}\}$, $Y^2_{n^1+1} \in \{C_{k_1+1}, \ldots, C_{k_2}\}$ and $Y^3_{n^1+1} \in \{C_{k_2+1}, \ldots, C_K\}$. Let $\hat{q}$, $\underline{q}$ and $\overline{q}$ denote the proportions of correctly classified future observations (out of $N = 10,000$) using equations (10), (11) and (12) for selecting the optimal thresholds k_1 and k_2, respectively. The results are summarised in Table 2.

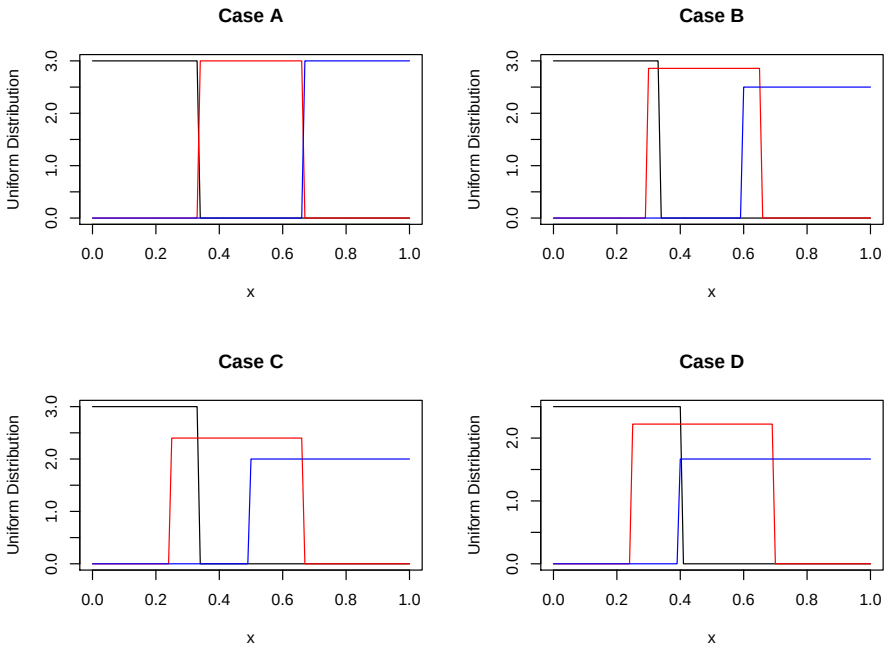

Fig. 1. Uniform distributions: Y^1 (black), Y^2 (red) and Y^3 (blue)

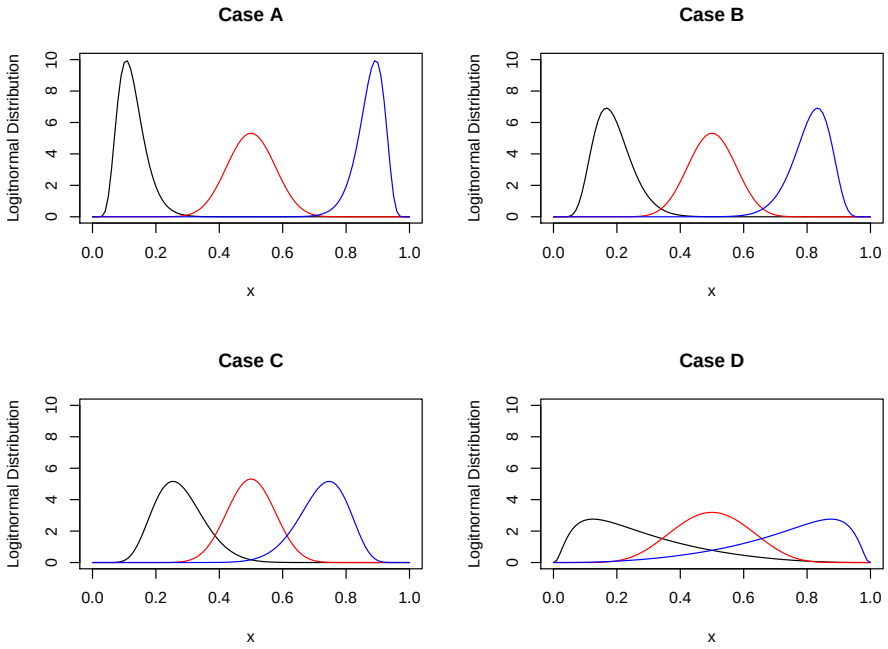

Fig. 2. Logit-Normal distributions: Y^1 (black), Y^2 (red) and Y^3 (blue)

Table 2. Simulation study

Cases	K	$n^1 = n^2 = n^3 = 100$						$n^1 = 180, n^2 = 100, n^3 = 10$					
		Uniform			Logit-Normal			Uniform			Logit-Normal		
		$\underline{q}$	$\hat{q}$	$\overline{q}$	$\underline{q}$	$\hat{q}$	$\overline{q}$	$\underline{q}$	$\hat{q}$	$\overline{q}$	$\underline{q}$	$\hat{q}$	$\overline{q}$
A	10	0.4876	0.4873	0.4873	0.8533	0.8533	0.8533	0.4543	0.4479	0.4543	0.8568	0.8559	0.8568
	20	0.7142	0.7142	0.7142	0.9342	0.9342	0.9342	0.7645	0.7633	0.7645	0.9505	0.9503	0.9505
B	10	0.3567	0.3566	0.3566	0.5421	0.5421	0.5421	0.3902	0.3849	0.3902	0.5245	0.5117	0.5245
	20	0.6297	0.6292	0.6293	0.8462	0.8462	0.8462	0.6148	0.6100	0.6148	0.8658	0.8643	0.8658
C	10	0.3295	0.3280	0.3285	0.3226	0.3226	0.3226	0.3422	0.3288	0.3422	0.3478	0.3450	0.3478
	20	0.4575	0.4565	0.4570	0.5693	0.5680	0.5691	0.4565	0.4460	0.4565	0.5951	0.5881	0.5951
D	10	0.2913	0.2911	0.2912	0.2671	0.2671	0.2671	0.2788	0.2653	0.2788	0.2752	0.2686	0.2752
	20	0.3152	0.3151	0.3154	0.3643	0.3639	0.3639	0.2990	0.2872	0.2990	0.3525	0.3483	0.3525
E	10	0.0159	0.0157	0.0160	0.0109	0.0107	0.0113	0.0173	0.0154	0.0173	0.0130	0.0119	0.0130
	20	0.0168	0.0163	0.0167	0.0116	0.0115	0.0121	0.0165	0.0160	0.0165	0.0137	0.0124	0.0137

For the unbalanced sample size scenario ($n^1 = 180, n^2 = 100, n^3 = 10$) the number of correctly classified future observations are higher for our NPI method compared to the empirical one, while for the balanced sample size scenario ($n^1 = n^2 = n^3 = 100$) the proposed method performs as good as the empirical one if not better. The equal performance related to the fact in most cases (in particular for small K) the proposed method and the empirical methods return the same optimal thresholds. There are also cases where the lower performs better than the upper and the other way around, but overall, the proposed method outperforms the empirical method in particular in case of unbalanced sample sizes. We also notice that all methods perform better for $K = 20$ than for $K = 10$, as for fixed $G = 3$ this may allow the methods to select more accurately the optimal thresholds k_1 and k_2 out of 20 classes rather than 10. Unsurprisingly for Case E, where the data are simulated form the same distribution, the numbers of correctly classified future observations are quite small.

4. Examples

Example 1

In this example we use the data set from Irwin and Irwin (2013) to compare the Organization for Economic Cooperation and Development (OECD) ratings made in early 2002 with a country's recourse to the International Monetary Fund (IMF) during the following nine years. The aim is to assess the discriminatory power of the OECD's country risk ratings in order to predict whether a country will have a program with the IMF, which is often used as an indicator of financial distress. Table 3 summarises the OECD risk classifications for 161 countries, 82 of which had recourse to an IMF program, and 79 of which did not have recourse to an IMF program. The OECD classifies countries on an eight-point scale from 0 (least risky) to 7 (most risky). Irwin and Irwin (2013) compared the Cumulative Accuracy Profile (CAP) accuracy ratio and the area under the ROC curves for the given data set. They showed that ROC analysis has several merits over the CAP curve, therefore the ROC curve was preferable by the authors. In this paper, we have compared the proposed method with the empirical area under the ROC curve.

Table 3. OECD Risk Rating and IMF Program status, Example 1.

	OECD Risk Rating							
IMF Program	0	1	2	3	4	5	6	7
Yes	3	0	1	2	5	8	13	50
No	21	2	12	14	8	4	5	13
Total	24	2	13	16	13	12	18	63

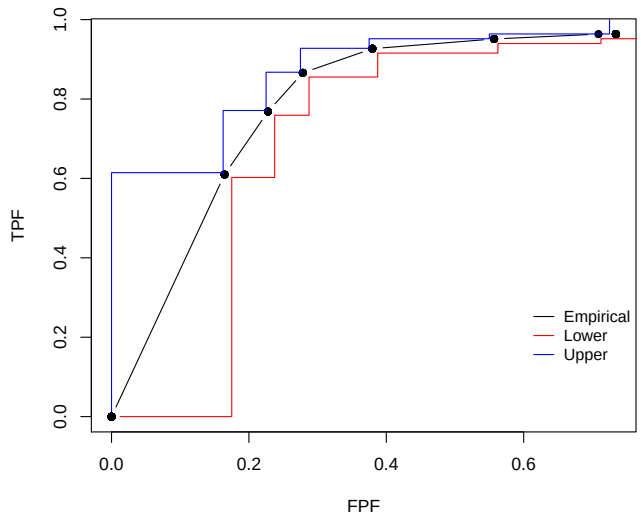

Fig. 3. The lower, empirical and upper ROC curves of OECD ratings as predictors of borrowing from the IMF in the following nine years, Example 1.

So in this example we have a two-group classification problem, so the ROC curve and its corresponding AUC are used, described in Elkhafifi and Coolen (2012), to asses the OECD ability to predict whether a country will have a program with the IMF. The ROC curve is defined as the combination of False Positive Fraction (FPF) and True Positive Fraction (TPF) over all values of the threshold k in $\{1, \ldots, K\}$, where $\text{FPF}(k) = P(Y^1 \in \{C_k, \ldots, C_K\})$ and $\text{TPF}(k) = P(Y^2 \in \{C_k, \ldots, C_K\})$.

Figure 3 shows the lower, empirical and upper ROC curves of OECD ratings as predictors of borrowing from the IMF in the following nine years. The corresponding areas under the curves are $\underline{AUC} = 0.7360$, $\widehat{AUC} = 0.8231$ and $\overline{AUC} = 0.8944$, respectively. These values show that OECD ratings has a good discriminatory power as a predictor of borrowing from the IMF. From predictive perspective, the large values of the lower and upper AUC show a strong evidence for a correct ordering of the two future observations. In classical method (e.g. the empirical AUC) one often performs a hypothesis testing to test the significant of the estimated AUC (e.g. $H_0 : AUC = 0.5$). In NPI context we do not perform any hypothesis testing, instead we see whether the value 0.5 is between the NPI lower and upper AUC, if this is the case we say that we have no or weak evidence

Table 4. Youden's indices, Example 1.

k	0	1	2	3	4	5	6	7
$\underline{J}(k)$	0	0.2143	0.2393	0.3773	0.5282	0.5679	0.5215	0.4274
$\hat{J}(k)$	0	0.2292	0.2546	0.3943	0.5471	0.5874	0.5404	0.4452
$\overline{J}(k)$	0	0.2389	0.2639	0.4018	0.5527	0.5925	0.5461	0.4520

that the future observations would be correctly ordered.

In addition, our method also attractive if one wants to compare two ROC curves, e.g. we have a further program in addition to the IMF program, let us refer to these programs as A and B. Then, instead of performing hypothesis testing as in the classical method to test the null hypothesis that $AUC_A = AUC_B$ we compare their lower and upper AUCs as follows. We say that we have a strong indication that classifier A is better than classifier B if the $\underline{AUC}_A = \overline{AUC}_B$. And we say that we have a weak evidence that classifier A is better than classifier B if $\underline{AUC}_A = \underline{AUC}_B$ and $\overline{AUC}_A = \overline{AUC}_B$.

Table 4 presents the values of Youden's index $\widehat{J}(k)$ for the empirical ROC curve together with Youden's indices corresponding to the NPI lower and upper ROC curves, $\underline{J}(k)$ and $\overline{J}(k)$, respectively. These indices are all maximal for $k = 5$, leading to the optimal OECD ratings being such that an outcome ratings of 5 to 7 indicates the country will have a program with the IMF, while an outcome ratings of 4 or less indicates the country will not have a program with the IMF.

Example 2

In this example, we consider a loan data set from a small Greek bank, a slightly different version of the data set has been used by Xanthopoulos and Nakas (2007) to introduce the empirical ROC surface for loan data. The bank used a credit rating system that assigns ratings from 1 to 8, where rates 7 and 8 are forbidden from getting loans. Therefore, the former two ratings are excluded from the analysis, and due to bank's size the remaining six ratings have been appropriately combined into 4 credit ratings, from 1 to 4, where obligors with rates 1 are considered to have excellent creditworthiness, while obligors with rates 4 are considered to have poor creditworthiness, and one would expect difficulties regarding their payment behaviour. Obligors are classified into four groups according to their delinquency status, i.e. whether or not there is a delay of payment on the last day of the year under consideration. Thus, they have been classified into group A if there is no delay of payment, into group B if the maximum delay is between 1 and 90 days, into group C if the delay is between 90 and 180 days, and into group D if they delay the payment for over 180 days.

First let us consider the following two-class scenarios: $A < (B + C + D)$, $(A + B) < (C + D)$, $(A + B + C) < D$. The results are summarised in Table 5. We can see from this table that this credit rating system can discriminate well between the first three groups combined and the last group. On the other hand, the credit rating system does not perform very well in the other two cases, as the areas under the empirical ROC curves are close to 0.5, while the areas under the lower NPI ROC curves are very small.

Table 5. Areas under the ROC curves, Example 2.

	$\underline{AUC}$	$\widehat{AUC}$	$\overline{AUC}$
$A < (B + C + D)$	0.2745	0.5192	0.7638
$(A + B) < (C + D)$	0.2875	0.5263	0.7649
$(A + B + C) < D$	0.4492	0.6363	0.8205

Table 6. Volumes under the ROC surfaces, Example 2.

	$\underline{VUS}^L$	$\underline{VUS}$	$\underline{VUS}^U$	$\widehat{VUS}$	$\overline{VUS}^L$	$\overline{VUS}$	$\overline{VUS}^U$
$(A + B) < C < D$	0.0644	0.0644	0.0648	0.2521	0.5695	0.5719	0.5722
$A < (B + C) < D$	0.0657	0.0657	0.0659	0.2536	0.5734	0.5751	0.5753
$A < B < (C + D)$	0.0221	0.0222	0.0224	0.1850	0.5233	0.5280	0.5288

Second, we use the three-group ROC methodology in order to assess the discrimination ability of this credit rating system, considering the three following cases: $(A+B) < C < D$, $A < (B + C) < D$, and $A < B < (C + D)$, where "+" means that the associated groups are combined. The results are given in Table 6. From this table we can see that $\underline{VUS}^L < \underline{VUS} < \underline{VUS}^U < \widehat{VUS} < \overline{VUS}^L < \overline{VUS} < \overline{VUS}^U$. For the case $A < (B + C) < D$, the NPI lower (upper) bound for the lower (upper) ROC surface is plotted in Figure 4 (Figure 5).

For the case where we compare the four groups $A < B < C < D$, using the formulas given in the appendix, the empirical hypervolume under the ROC hypersurface is $\widehat{VUHS} = 0.0646$, the NPI lower and upper bounds for the lower hypervolume under ROC hypersurface are $\underline{VUHS}^L = 0.00032$, $\underline{VUHS}^U = 0.00033$, the NPI lower and upper bounds for the upper hypervolume under the ROC hypersurface are $\overline{VUHS}^L = 0.37337$, $\overline{VUHS}^U = 0.37940$. One can see that the lower (upper) bounds are very close to each other, therefore they provide a good approximation for the exact lower (upper) ROC hypersurface. We can also see that the empirical hypervolume is between the NPI lower and upper hypervolumes, however it is much closer to the lower hypervolume than to the upper. The small value of the empirical VUHS, $\widehat{VUHS} = 0.0646$, indicates poor performance of the bank's adopted credit rating system, yet it is somewhat better than a random classier (1/24=0.04167). However, from predictive perspective the lower VUHS indicates that we have very little evidence that the four next borrowers, one from each group, would all be correctly ordered but the upper VUHS shows that we have large imprecision, and hence there is no strong evidence against the possibility of such a correct ordering. Although this is a considerable data set, several categories have only very few observations leading to substantial imprecision.

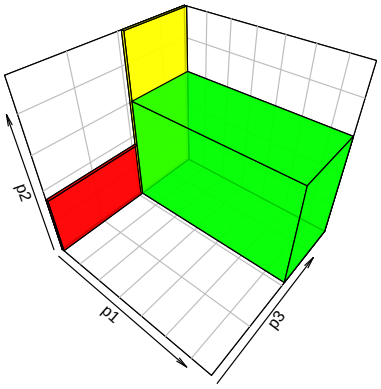

Fig. 4. The lower bound for the lower ROC surface, Example 2.

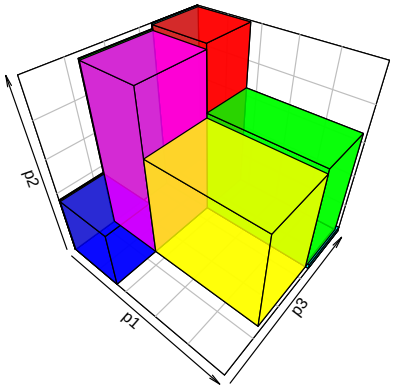

Fig. 5. The upper bound for the upper ROC surface, Example 2.

5. Concluding remarks

In this paper we introduced NPI for ROC analysis within a banking context, which includes NPI lower and upper bounds for the hypervolumes under the lower and upper ROC hypersurfaces. The presented work can be extended in many ways. For example, Coolen-Maturi (2017a) introduced NPI for combining real-valued diagnostic tests taking into account the problem of limits of detection, that is when data are unobservable above or below certain limits. Such methods can also be used in case of many ordinal categories, if one cannot get perfect observations. Furthermore, NPI has been presented for direct selection of the optimal thresholds of a diagnostic test based on multiple future observations, which shows some promising results compared to the use of the Youden's index for three-group classification problems (Coolen-Maturi et al., 2018). Extending that for ordinal data and for more than three groups is an interesting topic.

Acknowledgement

We are grateful to Dr. Stelios Xanthopoulos and Dr. Christos Nakas for providing the data used in Example 2. The authors thank the Associate Editor and two reviewers for supportive and constructive comments and suggestions.

Appendix

In this section we write out the formulas presented in Section 2 for the special case when $G = 4$ as follows:

$$\begin{aligned}
\widehat{VUHS} = \frac{1}{\prod_{g=1}^{4} n^g} \Bigg[& \sum_{i_1=1}^{K-3} \sum_{i_2=i_1+1}^{K-2} \sum_{i_3=i_2+1}^{K-1} \sum_{i_4=i_3+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 + \frac{1}{2} \sum_{i_1=1}^{K-2} \sum_{i_2=i_1+1}^{K-1} \sum_{i_3=i_2+1}^{K} n_{i_1}^1 n_{i_1}^2 n_{i_2}^3 n_{i_3}^4 \\
& + \frac{1}{2} \sum_{i_1=1}^{K-2} \sum_{i_2=i_1+1}^{K-1} \sum_{i_3=i_2+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_2}^3 n_{i_3}^4 + \frac{1}{2} \sum_{i_1=1}^{K-2} \sum_{i_2=i_1+1}^{K-1} \sum_{i_3=i_2+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_3}^4 \\
& + \frac{1}{6} \sum_{i_1=1}^{K-1} \sum_{i_2=i_1+1}^{K} n_{i_1}^1 n_{i_1}^2 n_{i_1}^3 n_{i_2}^4 + \frac{1}{6} \sum_{i_1=1}^{K-1} \sum_{i_2=i_1+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_2}^3 n_{i_2}^4 + \frac{1}{24} \sum_{i_1=1}^{K} n_{i_1}^1 n_{i_1}^2 n_{i_1}^3 n_{i_1}^4 \Bigg] \\
\underline{VUHS}^L = \frac{1}{\prod_{g=1}^{4} (n^g + 1)} \Bigg[& \sum_{i_1=1}^{K-3} \sum_{i_2=i_1+1}^{K-2} \sum_{i_3=i_2+1}^{K-1} \sum_{i_4=i_3+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 \\
& - \sum_{i_1=1}^{K-3} \sum_{i_3=i_1+2}^{K-1} \sum_{i_4=i_3+1}^{K} n_{i_1}^1 n_{i_3}^3 n_{i_4}^4 - \sum_{i_1=1}^{K-3} \sum_{i_2=i_1+1}^{K-2} \sum_{i_4=i_2+2}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_4}^4 + \sum_{i_1=1}^{K-3} \sum_{i_4=i_1+3}^{K} n_{i_1}^1 n_{i_4}^4 \Bigg] \\
\underline{VUHS}^U = \frac{1}{\prod_{g=1}^{4} (n^g + 1)} & \sum_{i_1=1}^{K-3} \sum_{i_2=i_1+1}^{K-2} \sum_{i_3=i_2+1}^{K-1} \sum_{i_4=i_3+1}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_4}^4
\end{aligned}$$

$$\overline{VUHS}^L = \frac{1}{\prod_{g=1}^{4}(n^g+1)}\left[\sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K}\sum_{i_3=i_2}^{K}\sum_{i_4=i_3}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 + \sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K}\sum_{i_3=i_2}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 + \sum_{i_2=1}^{K}\sum_{i_3=i_2}^{K}\sum_{i_4=i_3}^{K} n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 + \sum_{i_2=1}^{K}\sum_{i_3=i_2}^{K} n_{i_2}^2 n_{i_3}^3\right]$$

$$\begin{aligned}\overline{VUHS}^U = \frac{1}{\prod_{g=1}^{4}(n^g+1)}\Bigg[&\sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K}\sum_{i_3=i_2}^{K}\sum_{i_4=i_3}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 + \sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K}\sum_{i_3=i_2}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_3}^3 \\ &+ \sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K}\sum_{i_4=i_2}^{K} n_{i_1}^1 n_{i_2}^2 n_{i_4}^4 + \sum_{i_1=1}^{K}\sum_{i_3=i_1}^{K}\sum_{i_4=i_3}^{K} n_{i_1}^1 n_{i_3}^3 n_{i_4}^4 + \sum_{i_2=1}^{K}\sum_{i_3=i_2}^{K}\sum_{i_4=i_3}^{K} n_{i_2}^2 n_{i_3}^3 n_{i_4}^4 \\ &+ \sum_{i_1=1}^{K}\sum_{i_2=i_1}^{K} n_{i_1}^1 n_{i_2}^2 + \sum_{i_1=1}^{K}\sum_{i_3=i_1}^{K} n_{i_1}^1 n_{i_3}^3 + \sum_{i_1=1}^{K}\sum_{i_4=i_1}^{K} n_{i_1}^1 n_{i_4}^4 + \sum_{i_2=1}^{K}\sum_{i_3=i_2}^{K} n_{i_2}^2 n_{i_3}^3 \\ &+ \sum_{i_2=1}^{K}\sum_{i_4=i_2}^{K} n_{i_2}^2 n_{i_4}^4 + \sum_{i_3=1}^{K}\sum_{i_4=i_3}^{K} n_{i_3}^3 n_{i_4}^4 + n^1 + n^2 + n^3 + n^4 + 1\Bigg].\end{aligned}$$

References

Augustin, T. and Coolen, F. P. A. (2004) Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, **124**, 251–272.

Baker, R., Coolen-Maturi, T. and Coolen, F. P. A. (2017) Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, **44**, 1333–1349.

Coolen, F. P. A. (2011) Nonparametric predictive inference. In *International Encyclopedia of Statistical Science* (ed. M. Lovric), 968–970. Springer.

Coolen-Maturi, T. (2017a) Predictive inference for best linear combination of biomarkers subject to limits of detection. *Statistics in Medicine*, **36**, 2844–2874.

Coolen-Maturi, T. (2017b) Three-group ROC predictive analysis for ordinal outcomes. *Communications in Statistics: Theory and Methods*, **46**, 9476–9493.

Coolen-Maturi, T., Coolen, F. P. A. and Alabdulhadi, M. (2018) Nonparametric predictive inference for diagnostic test thresholds. In submission.

Coolen-Maturi, T., Coolen-Schrijner, P. and Coolen, F. P. A. (2012a) Nonparametric predictive inference for binary diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 665–680.

Coolen-Maturi, T., Coolen-Schrijner, P. and Coolen, F. P. A. (2012b) Nonparametric predictive inference for diagnostic accuracy. *Journal of Statistical Planning and Inference*, **142**, 1141–1150.

Coolen-Maturi, T., Elkhafifi, F. F. and Coolen, F. P. (2014) Three-group ROC analysis: A nonparametric predictive approach. *Computational Statistics & Data Analysis*, **78**, 69–81.

Crook, J. N., Edelman, B., D. and Thomas, L. C. (2007) Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, **183**, 1447–1465.

De Finetti, B. (1974) *Theory of Probability.* London: Wiley.

Elkhafifi, F. F. and Coolen, F. P. A. (2012) Nonparametric predictive inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 681–697.

Hand, D. J. and Henley, W. E. (1997) Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A*, **160**, 523–541.

He, T., Coolen, F. P. A. and Coolen-Maturi, T. (2018) Nonparametric predictive inference for european option pricing based on the binomial tree model. *Journal of the Operational Research Society*, accepted.

Hill, B. M. (1968) Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, **63**, 677–691.

Irwin, R. J. and Irwin, T. C. (2013) Appraising credit ratings: Does the cap fit better than the ROC? *International Journal of Finance & Economics*, **18**, 396–408.

Nakas, C. T. and Yiannoutsos, C. T. (2004) Ordered multiple-class ROC analysis with continuous measurements. *Statistics in Medicine*, **23**, 3437–3449.

Nakas, C. T., Alonzo, T. A. and Yiannoutsos, C. T. (2010) Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index. *Statistics in Medicine*, **29**, 2946–2955.

Pepe, M. S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford: Oxford University Press.

Wutzler, T. (2018). logitnorm: Functions for the Logitnormal Distribution. R package version 0.8.37. https://CRAN.R-project.org/package=logitnorm.

Xanthopoulos, S. Z. and Nakas, C. T. (2007) A generalized ROC approach for the validation of credit rating systems and scorecards. *Journal of Risk Finance*, **8**, 481–488.

Youden, W. J. (1950) Index for rating diagnostic tests. *Cancer*, **3**, 32–35.