Alma Mater Studiorum
Università di Bologna

DOTTORATO DI RICERCA IN
SCIENZE STATISTICHE

CICLO XXX

Settore Concorsuale: 13/D1
Settore Scientifico Disciplinare: SECS-S/01

Ordinal data supervised classification with Quantile-based and other classifiers

Presentata da: Lorenzo Mancini

Coordinatore Dottorato:
Prof.ssa Alessandra Luati

Supervisore:
Prof.ssa Cinzia Viroli
Co-Supervisore:
Prof. Christian Hennig

Esame finale anno 2018

Alma Mater Studiorum
University of Bologna

PHD DEGREE IN
STATISTICAL SCIENCES

CYCLE XXX

Competition field: 13/D1
Academic discipline: SECS-S/01

Ordinal data supervised classification with Quantile-based and other classifiers

Presented by: Lorenzo Mancini

Ph.D. Director:
Prof. Alessandra Luati

Supervisor:
Prof. Cinzia Viroli
Co-Supervisor:
Prof. Christian Hennig

Final exam year 2018

Abstract

The aim of this research project is to propose a new method for supervised classification problems where the input features are ordinal. Ordinal data are preponderant in many research fields. They directly arise when the observations fall into separate distinct but ordered categories and they are very common in surveys where answers are listed as Likert scales. Typically, they are coded as equally spaced values and sometimes they are analyzed as numerical values. These choices may not necessarily correspond to the real distribution of the data.

The objectives of the study have been accomplished according to several steps. The first phase consisted of an exhaustive analysis of the state of art of the statistical literature with the aim of identifying the various approaches to ordinal data analysis, the related limitations, and possible advantages. We have then proposed to operate in the framework of Generalized Linear Latent Variable Models (GLLVM), considering the response function approach with a single latent variable Beta distributed. Our scope in using this method is to shift from a set of ordinal features to a single continuous feature, which well adapt the data, in order to directly apply the standard classification methods.

A dedicated EM algorithm has been developed on the basis of this theoretical framework using the statistical software **R**.

Finally, we have compared our approach with several scoring methods through a wide simulation study. The scoring methods that we have considered in the simulation study are: the raw scores, the ridits, the blom scores, the normal median scores and the conditional mean scores. These methods, although have a long history in literature, have never been used for classification purpose.

In addition we present an example of the application of the proposed approach to real world business data problem.

Sommario

Il lavoro di ricerca ha l'obiettivo di individuare una metodologia statistica per la classificazione supervisionata di unità statistiche misurate da un insieme di variabili ordinali. In generale, si parla di variabili ordinali ogniqualvolta il carattere assume stati discreti ma ordinabili. Questo tipo di dati è molto diffuso in diverse aree di ricerca e, in particolare, è molto comune nei sondaggi, dove le categorie di risposta sono elencate tramite scale Likert. Tipicamente, le categorie associate a queste variabili sono codificate attraverso apposite etichette. Le etichette corrispondono solitamente a valori numerici progressivi ed equi-distanziati che riflettono l'ordine delle categorie. In fase di analisi non è però appropriato trattare questi dati come valori numerici reali, in quanto, così facendo, si andrebbe ad introdurre una distanza tra categorie che potrebbe non corrispondere a quella effettiva.

Il progetto di ricerca si articola in diverse fasi. Inizialmente, viene effettuata un'analisi esaustiva dello stato dell'arte della letteratura, per identificare i vari approcci all'analisi dei dati ordinali, valutandone i limiti e i vantaggi.

Successivamente, sulla base dei risultati di questa analisi, viene proposto un metodo basato sull'approccio response function, nel contesto dei modelli generalizzati a variabili latenti. A differenza del metodo classico, che prevede variabili latenti normalmente distribuite, la nuova metodologia proposta considera una singola variabile latente con distribuzione Beta, poiché fornisce specifici vantaggi in termini di efficienza computazionale e di adattamento ai dati. L'obiettivo è, sostanzialmente, di spostare il problema della classificazione da un insieme di variabili ordinali ad una singola variabile continua, in modo da applicare i metodi di classificazione standard.

Sulla base di questo quadro teorico di riferimento è stato sviluppato un algoritmo EM, utilizzando il software statistico **R**.

Infine, l'approccio proposto è confrontato, attraverso un ampio studio di simulazione, con diversi metodi di scoring, in particolare: raw scores, ridits, blom scores, normal median scores e conditional mean scores. Questi metodi non sono mai stati usati per scopi di classificazione, sebbene abbiano una lunga tradizione nella letteratura.

Si presenta, in aggiunta, un'applicazione del metodo discusso ad un problema di classificazione su dati reali.

Acknowledgements

First and foremost I would like to thank my advisor Prof. Cinzia Viroli for her constant support of my Ph.D study and research, for her contributions of time, motivation, enthusiasm, gentle encouragement and faith in me during the dissertation process.

Her guidance was invaluable and helped me in all the time of research and writing of this thesis.

Besides my advisor, I would like to thank my co-supervisors Prof. Christian Hennig for inviting me at the University College of London. This has been a great opportunity and helped me to grown as researcher. I thank him for the academic support, for his insightful comments and for all the illuminating discussions we had.

Lastly, I would like to thank my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The context

It is very common, in many different areas of statistical research, to deal with data measured on the ordinal scale. These kind of data are particularly preponderant in behavioural, political, educational, psychological and social sciences and they generally arise in all the contexts where it is not possible to obtain a finer representation of the statistical unit attribute due to the nature of the observed phenomenon or the availability of measuring instruments.
A typical example of such data, coming from the psychometric field, is the Likert scale, which is a technique consisting in developing a number of statements (items) that express a positive or negative attitude to a specific aspect. Respondents are asked to express their degree of agreement or disagreement with respect to a specific statement, usually on a 5-points or 7-points scale such as:

"strongly disagree" "disagree" "no opinion" "agree" "strongly agree"

The analysis of ordered response variables has become increasingly important in the last decades and many specific approaches has been proposed. Anyone of these approaches have to face with the unique challenges that ordinal features present: on one hand they differ from nominal data as order information is present and it has to be considered in the analysis, on the other hand they also differ from interval scaled data as they do not include the notion of distance between categories. Despite the vast body of literature on this topic, only few methods address the specific task of classifying ordinal

data, which is the purpose of the present study. Specifically, the aim is to find a suitable way for classifying these data in the supervised context, with particular reference to the quantile-based classifier, proposed by Hennig & Viroli (2016a), which will be described in detail in the following chapters.
Before proceeding further, it seems reasonable to define the notion of ordinal data since, although this seems quite intuitive, it has been the subject of several discussions over the years.

1.2 Ordinal data

In order to define what the ordinal scale of measure is, it is necessary to start from the early stage of measurement theory. Measurement theory is a branch of applied statistic that attempts to describe and evaluate the quality, the usefulness and the meaningfulness of measurements.
In the classical definition, measurement is the expression of some characteristic through a real number times a unit (e.g., metres, grams). Thus, following this definition, ordinal variables are not strictly considered as a measurement as no unit of measurement is defined. However, the notion of ordinal scale of measurement has been reintroduced in the early 1940's by Stevens (Stevens *et al.*, 1946), which has reformulated the concept of measurement in a more general way as the assignment of numerals to objects or events according to rules. Different sets of rules lead to different kinds of scales and different kinds of measurements.
Stevens first coined the terms nominal, ordinal, interval and ratio to describe a hierarchy of measurement scales based on invariance of their meaning under different classes of transformations.
Table 1.1 from Steven's paper summarizes each scale by listing: the basic empirical operations associated with it (column 2); the mathematical group structure, i.e. the mathematical transformations which leave the scale-form invariant (column 3) and the statistics that it is possible to use on the relative scale type of data that preserve the invariance under the transformations in the third column (column 4).
It needs to be pointed out that every column in the table is cumulative in the sense that going from nominal to ratio scale the possible empirical operations for a particular scale must be added to all those operations preceding it. The same is true for the Mathematical group structure (i.e. each mathematical group is contained in the group immediately above it) and for the permissible

statistics associated to the scale (thus, if we are in the interval scale we can perform mean and standard deviation as well as median and mode).
According to Stevens, only the interval scaled variables fall within the classical definition of measurement scale.

Table 1.1: *Scales of Measurement.*

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	Permutation group $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	Isotonic group $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	General linear group	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	Similarity group $x' = ax$	Coefficient of variation

Once a consistent set of rules under which numerals are assigned to attributes are defined, one should be able to understand the kind of measurement. This set of rules correspond to all the isomorphisms (the set of appropriate transformation) of the numerical attribute. Some authors (see Kampen & Swyngedouw, 2000) pointed out that, as we do not know the "real" value of the attribute before actually measuring it, we are not able to say whether a particular transformation is appropriate or not and this is a serious limitation in the definition of measurement scale.
Although the critics moved over the years to this theory, we decide to proceed with Steven's ordinal scale definition as it is the most adopted in practice (Agresti, 2003), both in statistical and non-statistical fields.
According to the definition of Stevens, the ordinal scale arises from the operation of rank-ordering, i.e. assigning to the statistical units the corresponding ranks. The ordinal scale has the isotonic or order preserving structure.
So, for an ordinal variable, say X, it is assumed that:

1. $X \in \{x_1, \ldots, x_k\}$, where $x_i \in \mathbb{R}, i = (1, 2, \ldots, k)$ and k is the number of exclusive and exhaustive categories.

2. The categories satisfy $x_1 < \ldots < x_k$.

1.3 Approaches to ordinal data analysis

The appropriateness and the meaningfulness of methods for dealing with ordinal scale of measurement has been the subject of considerable controversy for several years and this controversy is still ongoing. We present the issue from a conceptual point of view, tracing the basic assumptions of each approach of handling ordinal variables and illustrating the motivations and the associated issues.
From the literature three major approaches emerge for treating ordinal categorical data. In differentiating them, we follow the nomenclature used by Kampen & Swyngedouw (2000) as it is of immediate comprehension:

- Parametric approach
- Non-parametric approach
- Underlying variable approach

1.3.1 The parametric approach

The parametric approach consists in replacing the categories with arbitrary numerical values and proceeding in the analysis by using the classical parametric inference methods such as ANOVA or OLS (directly on the alternative scores or after arithmetic synthesis of them).
Despite this "unsophisticated" approach has been strongly criticized by purists, both from a theoretical and practical point of view, it is still used due to its simplicity and immediacy in applications.
Stevens himself, in describing the ordinal scale, provides a justification for the use of ordinary statistics in this context:

> [...] for this "illegal" statisticizing there can be invoked a kind of pragmatic sanction: in numerous instances it leads to fruitful results.

A great support to this approach is also due to the work of Labovitz, where simulations were run to demonstrate that the use of ordinary statistics with ordinal data does not lead to large errors (Labovitz, 1970 and Labovitz, 1971). However, in a subsequent work, O'Brien showed that the underlying latent distribution and the number of categories have an effect upon the size of the errors (O'Brien, 1979).

The fundamental critic to this approach is that the original scale is an ordinal scale, without the concept of distance. Once we introduce numbers, we are implicitly defining arbitrary distances between categories, which exaggerates the information provided by the data. As the categories reflects only ordinality, the differences between these numeric codes have no meaning. For example, Clogg asserts that on a three-point happiness scale the distance between "not too happy" and "pretty happy" categories is about three times greater than the distance between "pretty happy" and "very happy" responses (Clogg, 1982).

> *If the responses are coded 0,1,2,3 or 4, a linear regression would treat the difference between a 4 and a 3 in the same way as a difference between a 3 and a 2, while in fact they are only ranking.* (Greene, 1993)

Although sometimes this approach may be useful and numerical values may well approximates "reasonable" continuous measurement for a first descriptive analysis or in evaluating the effects of covariates on a response variable, one should proceed very carefully with the conclusions drawn from the analysis of these data. Since the numeric codes are assigned at the limit of arbitrariness, it may happen that, changing the numerical attribute of one or more categories (even leaving the ordinality unchanged) will lead to different results of the analysis and this also implies that the analysis may be controversial.

In order to give a unique interpretation to the analysis on these data a suggested practical approach could be to agree with a unique coding upon which to base all the analyses. As a matter of fact, a widely popular choice is to assign the rank ordering of the categories as numerical value. However, nothing guarantees that this is the best way to proceed. Furthermore, if we were allowed to use such scale as interval then there would be no point at all in distinguishing between the two scales of measurement.

With reference to the work of Agresti we report here a practical example of the limits that occur in the specific case of using a standard linear regression with an ordinal response variable (Agresti, 2003). Though these models can be used effectively to determine the effect of covariates on an ordinal response variable they present strong limitations that, for the most part, can be extended to any other kind of analysis:

1. It is usually not clear what scores should be. As mentioned, different set of scores may lead to different results in the analysis.

2. If ordinal response variables come as the discretization of some continuous variable we do not take into account the measurement error introduced by replacing an interval or ratio scale with an ordinal one.

3. It is likely to obtain predicted values above the highest category or below the lowest.

4. The variability of the responses is nonconstant for categorical data.

5. Ordinary regression approach does not account for "ceiling effects" and "floor effects" due to the fixed number of categories.

Regarding the fifth point we present a practical example from Agresti (2003), where a standard linear regression is applied to simulated data. The data set was generated considering a continuous uniform covariate x and a binary covariate z (see Agresti, 2003 for details about data simulation). The ordered categorical response variable y was generated from a underlying normally distributed variable y^*. Figure 1.1 shows the observations in the dataset on y^* and x (left panel) and on y and x (right panel), where the data points are labelled according with the value of z. The plot also shows the OLS fit of both models.

As it is possible to notice, going from continuous to ordered response variable there is a very high probability for observations to fall in the lowest category of y when $x < 50$ and $z = 1$. This floor effect causes the need to include in the model a non necessary interaction term or a quadratic effect of x on $E(y)$.

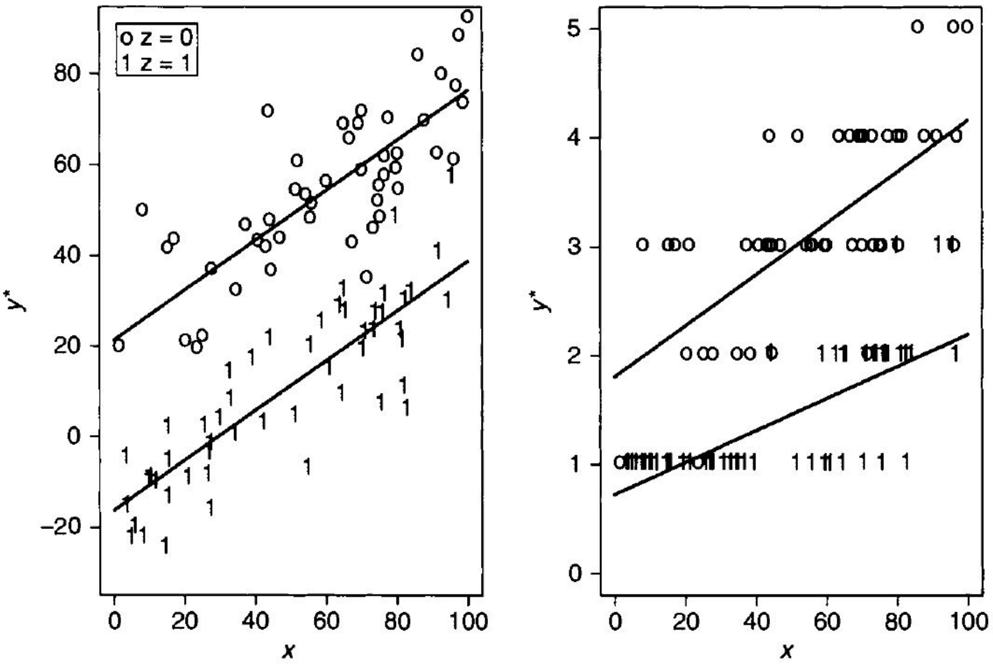

Figure 1.1: *Floor effect.*

1.3.2 The non-parametric approach

Advocates of non-parametric approach claim that the mathematical group structure of ordinal scale identified by Stevens prevents the use of models developed for interval scaled variables.
Thus, the analyses are restricted solely to methods that only use ordering information about the categories. No assumption are made on the distribution of the ordinal variable.
Examples are the methods based on ranks such as the Wilcoxon signed-rank test, used when comparing two related samples or repeated measurements on a single sample to assess whether their population mean ranks differ. A further example of non parametric approach is the proportional odds model (Agresti, 2003; McCullagh, 1980), as it only uses the ordering information in the categories of the response variable, thus it is not sensible to the numerical values assigned to the categories. In addition, we cite the CUB (**C**ovariates in **U**niform and shifted **B**inomial mixtures) models (Iannario & Piccolo, 2012), which have been developed in the last few years with the

aim of model the respondent's rating process, in terms of probability of responding in a specific category, through a weighted combination of feeling and uncertainty towards the item/object and BOS (**B**inary **O**rdinal **S**earch distribution) models, which assume that an ordinal variable is the result of a stochastic binary search algorithm within an ordered table, from 1 to the maximum possible category level (Biernacki & Jacques, 2016; Jacques & Biernacki, 2017).

These methods allow to avoid the problem of absence of an interval scaled measure and they proved to be fairly powerful. We explained that using integer scores and treating ordered response variables as if they were continuous may be questionable, specially if data have skew distributions. The use of model-based approaches, which avoid scoring, gives the opportunity to remove the arbitrariness consisting in assigning scores. However, strict adherence to operations that utilize only the ordering scales limits the scope of a useful methodology too severely.

1.3.3 The underlying variable approach

Because this terminology includes a wide range of possible approaches, we define the underlying variable approach in the broadest sense of the term as recording the ordinal categories so that parametric statistics can be applied. Unlike the parametric approach, here numeric values are assigned to categories in a meaningful way so that they meet as closely as possible some theoretical distributional assumptions.

Broadly speaking, this consists in assigning numbers to the categories that reflect the researcher's knowledge of an appropriate mathematical distances between the categories. So, rather than assigning arbitrary numeric values to the categories, we perform inference on parametric models for the latent variable, which is often more sensible.

Usually this is done by assuming a priori that the ordinal observed variables are the result of the discretization of an unmeasurable underlying variable. The ordinal scale comes from the categorization of an inherently continuous scale that is not possible to observe.

Objections often surrounds the assumption of an underlying continuous variable (see Kampen & Swyngedouw, 2000), since:

1. There is no way to prove that data actually comes from an underlying variable, often ordinal is the best one can do.

2. Even if the underlying variable exists, it is not possible to test any assumption required for a correct and meaningful use of inferential parametric statistics (e.g. normality assumption, homoschedasticity).

3. Given the first and second objections, it also becomes problematic to understand to what extent conclusions one may draw from data are valid or generalizable.

On the basis of whether an ordinal variable can or cannot be derived from a measurable (or not) underlying variable, Kampen and Swyngedouw made an useful distinction among five types of ordinal variables (Kampen & Swyngedouw, 2000). They also consider the case where an objective standard can be defined in order to calibrate a measurement instrument for the ordinal data (this is necessary to define a unit of measurement that is not dependent on the experimenter taking the measures).

- **Type I:** *The categorized metric variable with known thresholds.*
  It comes as a result of the categorization of a known measurable underlying variable.
  Example: classifying annual income selecting as threshold values 30.000€ and 50.000 € as "low"=< 30.000€, "middle"=30.000€ – 50.000€ and "high"=> 50.000€.

- **Type II:** *The categorized metric variable with unknown thresholds.*
  The underlying variable is measurable but classification cannot be done with reference to the units of this underlying variable.
  Example: classifying annual income in "low", "middle" and "high".

- **Type III:** *The categorized latent variable with unknown thresholds.*
  The underlying variable is not measurable.
  Example: psychiatrists classifying patients in having "low", "moderate" and "high" intelligence.

- **Type IV:** *The semi-standardized discrete variable with ordered categories.*
  Ordinal variable that cannot be conceived of having an underlying variable.
  Example: biologists classifying the young of intoxicated mice in "dead", "handicapped" and "sound".

- **Type V:** *The unstandardised discrete variable with ordered categories.* Ordinal variable that cannot be conceived of having an underlying variable and reference to an objective standard is difficult or impossible. Example: classifying the level of agreement with respect to a specific statement.

Only variable of type I have an objective standard while for variables of type III and IV standardization can be obtained by maximizing the agreement of experimenters taking the measures.
Kampen and Swyngedouw suggest to proceed by choosing a model that is the most appropriate with the variable type we deal with as uncalibrated measurements affects the validity of any method of analysis (Kampen & Swyngedouw, 2000).
Objections are made from authors who assert that preservation of order is all that is required and that any monotonic transformation of a set of numbers would do as well as any other, thus any attempts to scoring are illusory and one should just refer to non parametric approaches. If the assumption of a particular functional form for a latent distribution seems reasonable this does not imply scores to behave in a certain way and it does not lead in general to a specific distributional requirement.
Suppose we are in the case of type I variables. If the ordered categories are generated by choosing some feasible cut-points over the continuum scale then there is no theoretical justification that the obtained variables should reflect the properties of the reference continuous distribution. There is no relationship between the assumed symmetry of the underlying distribution of any assumed latent scale and the symmetry over the scores. If subjective unequal interval scaled scores are arbitrarily chosen they could be asymmetric even if the assumed underlying variable is symmetric. One possibility could be to test the ordered variable distributional form when this is required by model assumption (as normality). Procedures are available for testing such assumption as the PRELIS program suggested by Jöreskog (1990), which uses the extra structure that joint normality imposes.
However, except for cases where we deal with type IV and V variables the assumption of known distribution is often not unduly restrictive in many, if not most, practical applications. If there is an underlying continuum then there will be a population distribution on that continuum which will induce the ordered classes.
Advocates of this approach proceed with the reasonable conviction that, by

scoring the ordinal categorical data so that they reflects some prior information and treating them as interval scaled data, there is, potentially, less loss of information than simply applying statistics to categories.

An alternative not mentioned in previous approaches is to ignore both the order of the categories and any quantitative information and to treat the variable as nominal, using indicator variables. This approach is common in practice but, beyond the fact that it completely ignores the structure and information brought by the data, its applicability for the analysis of large data set is limited by the number of parameters introduced. As the number of variables and categories increase, the number of parameters involved in the model becomes enormous and performing the estimates become cumbersome. If the data are treated as interval scaled there are fewer coefficients to be estimated (and possibly more stability in the results).

In addition to what said until now, it has to be pointed out that the number of categories considered is of particular importance. Sometimes researchers work under the assumption that, with a sufficiently large number of categories, categorical data tend to be similar to continuous data, then classical statistical methodology may be applied directly.
Studies have examined the use of classical statistical analyses with ordinal data when the number of categories increase. Examples are Rhemtulla *et al.* (2012) and Beauducel & Herzberg (2006), which compare methods for estimating confirmatory factor analysis models with ordinal variables with different number of categories.
In the present study, however, the goal is to identify specific solutions to problematic cases, that is, in contexts where we have a narrow set of possible categories.

A large number of models for ordinal variables based on the presented approaches have been developed over time, each with its strengths and limitations. For the interpretation of these models it is necessary to meticulously look into the assumptions made in the models in order to be able to interpret the analysis outcome.

1.4 Thesis outline and objectives

In the present study we choose to proceed following the latent variable approach. Regarding the criticisms moved on the existence and the usefulness of a latent variable, we choose what could be called a "pragmatic" solution that allows us to operate free of these methodological limits. We move away from Platonic point of view that there is a world with "true values" with respect to which we should try to obtain the best possible approximation in order to perform the analysis.

Here the focus is not the "true model" but operating in a framework that approximates reality to a degree level sufficient for the practical purpose of supervised classification. This means that adequate predictions of the classes can justify the existence of an underlying variable and the use of parametric statistics, even if model assumptions are not fulfilled.

The objective is therefore to assess whether it is possible, through reasonable assignment of numerical values to the categories of the ordinal variables, to obtain good results in the context of supervised classification.

Methods for assigning numeric values to categories, commonly known as scoring methods, have a long history in literature but they have never been used for classification purposes.

Seven different scoring methods, which will be described in detail in the next chapter, have been considered in this research study.

We also propose a new methodology that we named Beta Response Function Approach (BRFA). Our proposal is based on the response function approach, developed in the context of Generalized Linear Latent Variable Models (GLLVM) and it allows to avoid the limits that emerge in the use of the considered scoring methods. The BRFA leads to fruitful results in the simulations performed.

The present work is structured as follows:

- In Chapter 2 the scoring methods considered in the simulations are discussed, together with their specific advantages and limitations.

- In Chapter 3 the new proposed method is presented along with the reasons that led us to formulate this innovative proposal.

- In Chapter 4 the classification methods adopted in the analyses are described.

- In Chapter 5 simulations results are presented and discussed.
- In Chapter 6 an example of the application of our method on a real world dataset is presented.
- In Chapter 7 conclusions and possible future research patterns are discussed.

Chapter 2

Scoring Methods

2.1 Previous studies

A scoring system is a systematic method for assigning numerical values to the variable categories. As mentioned in the previous chapter, choosing a particular set of scores does not guarantee that the assumptions of the chosen model are always verified. The central issue is the choice of the scoring scale and not whether scoring measures per se are appropriate. The study of scoring methods has a long story in the literature. An example is Yates (1948), which analysed data coming from a pilot inquire into the conditions in which school children do their homework. Data are reported below in a contingency table with both column and row ordered variables that can be regarded as having an underlying quantitative basis (Table 2.1). In this case, a common procedure for testing independence is to perform a χ^2 test. χ^2 test covers all forms of departure from proportionality and it is consequently insensitive to departures of a particular type.

Yates suggests to perform a traditional regression analysis to test for independence after appropriately assigning scores that for convenience are centred at 0 and are equally spaced. To the i^{th} row category the score $(2i - r - 1)/2$ is assigned and to the j^{th} column category the score $(2j - c - 1)/2$ is assigned. The χ^2 test becomes then a test of a zero regression coefficient from the model performed over the scores. Yates extended the analysis also to the case where just one variable in the contingency table is ordinal, thus re-conducing to the case of one-way analysis of variance.

Another example of analysis of data in the form of contingency tables, in which association is known to exist, can be found in the work of Fisher

(1940) (further developed by Williams, 1952), where scores are chosen to maximize the correlation between variables. These analyses have the advantage of giving tests of association more sensitive than the overall χ^2 test, and also of providing a practical interpretation of the category values.

A quite different procedure is shown by Snell (1964). In this work the scores are computed so that they can be used in the analysis of variance methods. The scoring system is determined so that it satisfies the two assumptions of residual deviations normally distributed and homogeneous residual variances. The categories are considered to come as a discretization of an assumed underlying continuous scale of measurement, which should be normally distributed. However, for reason of simplicity in computations, the distribution is assumed to be a logistic as the it agrees closely over most of its range with the normal curve. Thus, the scores are assigned by first detecting the threshold values x_i for $i = 1, \ldots, k$ (where k is the number of categories) by maximizing the likelihood and then computing the mid-points. Since the origin is arbitrary $x_1 = 0$ is assigned as the upper limit of the first category (with the lower limit set at minus infinity). For the first and the last categories scores are computed through an approximation algorithm. The scores are then applied to one-way analysis of variance.

Also Bollen (1989) has reported studies in factor analysis and structural equation models using a variety of scoring systems, including equally spaced integer scoring and polychoric induced mid-points.

Table 2.1: *Relation (in terms of numbers of children and percentages) between conditions under which homework was carried out, and the teacher's rating of the quality of that homework. Each scale is graded, "A" being the highest rating.*

Teacher's rating	Homework conditions					
	A	B	C	D	E	Total
A	141 (46%)	67 (46%)	114 (39%)	79 (44%)	39 (43%)	440 (43%)
B	131 (42%)	66 (45%)	143 (48%)	72 (40%)	35 (39%)	447 (44%)
C	36 (12%)	14 (9%)	38 (13%)	28 (16%)	16 (18%)	132 (13%)
Total	308 (100%)	147 (100%)	295 (100%)	179 (100%)	90 (100%)	1019 (100%)

2.2 The scoring methods

From the literature review several different methods of scoring have emerged. They are motivated by a variety of considerations but usually so that traditional statistical procedures could be adapted. The key aspect is the choice of appropriate relative distances between pairs of adjacent categories.
In the present study we have considered five types of scoring methods, among the main and most commonly used, for their attractiveness in our particular case of supervised classification.
Before examining in detail the various methods, we define:

- X, some ordinal categorical variable.
- k, the number of categories of X.
- $n_1, \ldots, n_k$ the frequencies of respondents in the categories, with $N = \sum_i^k n_i$.
- $p_1, \ldots, p_k$, the corresponding sample proportions, with $p_i = n_i/N$.

Following the notation in Brockett (1981) we denote the score associated with the i^{th} category as $s_i = h_k\left(i, p_1, \ldots, p_k\right)$, and $S = \{h_k\left(\bullet, \bullet, \ldots, \bullet\right)\}$ the scoring system determined by some scoring functions $h_k\left(\bullet, \bullet, \ldots, \bullet\right)$. The response probabilities $(p_1, \ldots, p_k)$ may depend upon some latent underlying distribution, the researcher knowledge about the phenomenon or to an empirical response distribution.
It appears obvious, as we do not refer to any particular property of the function upon which the scores are determined, that there is not a single "superior" approach to the treatment of ranked categorical data; it depends on the researcher's purposes and situation. Here we present five different scoring methods that will be used in the following simulations for classifications pourposes:

- Raw scores;
- Ridit scores;
- Normal median scores;
- Blom scores;
- Conditional mean scores.

2.2.1 Raw scores

We include raw scores as this particular scoring system is with no doubt the most widely used in many different fields. Raw scores emerges by replacing the categories with their the corresponding rank, thus non-negative integers between 1 and k. Raw scores are hardly to classify in one of the previous approaches for ordinal data: we are at the limit of the parametric approach in using this kind of scores because numerical values assigned to the categories may be directly applied into standard parametric inference methods. Also, we are at the limit of the non-parametric approach because the order information is used. There is no intention here to approximate any distributional form nor to utilize any prior knowledge about the data. Considering this, they still have rights to be included, in the broadest sense, in the context of underlying variable approach as equi-spaced scores may reflect the lack of knowledge about the distribution from which the data comes. They can be seen as coming from a discretization of a uniform latent distribution.
We proceed by including the raw scores in the following simulations considering them as a benchmark for the method we propose. Often for descriptive summaries it is more sensible to use fixed, equi-spaced scores such raw scores instead of scores based on data. Moreover, in some cases (see Fielding, 1993) they can compete with more specific scores.

2.2.2 Ridit scores

The ridit is a scoring system first introduced by Bross (1958). The first three letters of the term stays for “Relative to an Identified Distribution” and the suffix “it” was added by analogy with the probit and the logit names because ridits represents a type of transformation. The distribution to which the term refers is the one from which the $(p_1, \ldots, p_k)$ are observed. Thus, the crucial point in ridit analysis is the choice of the observed distribution upon which the computations are based. Among the various possible applications, the ridits are commonly used in epidemiology for analysis of ordinal data and also useful for analysis of questionnaires. The ridit value r_i for category i is defined as:

$$r_i = \frac{1}{2}\left(\pi_{i-1} + \pi_i\right)$$

where $\pi_i = \sum_{j \leq i} p_j$ is the cumulative sample proportion.
We report an example of ridit computation from Bross (1958) in Table 2.2.

The reference data set comes from the Cornell Automotive Crash Injury Research Program (ACIR) and reports the severity of the injury after a car accident (from “none” to “fatal”). Each column in the table represent a step in ridit computation.

Table 2.2: *Calculation of Ridits (Computing Form)*

	(1)	(2)	(3)	(4)	(5)
None	17	8.5	0	8.5	0.047
Minor	54	27	17	44	0.246
Moderate	60	30	71	101	0.564
Severe	19	9.5	131	140.5	0.785
Serious	9	4.5	150	154.5	0.863
Critical	6	3	159	162	0.905
Fatal	14	7	165	172	0.961
Total	179		179		

Column (1): The frequency distribution in the identified distribution (reference class).
Column (2): Half of the corresponding entry in Column (1).
Column (3): The cumulate of Column (1) (displaced one category downward).
Column (4): Column (2) + Column (3).
Column (5): The entries in Column (4) divided by grand total (ridits).

It is important to notice that ridit scores always vary between 0 and 1 and the mean ridit applied to the identified group will be identically 0.5. Since all scores are on the same range, categorical variables with different numbers of response categories become readily comparable. The characteristics of ridits make them suitable for the analysis of large questionnaires where model based approaches such log-linear models are limited by the large number of parameters.

The ridit analysis proposed by Bross has the purpose of comparing different groups of individuals with the reference group (i.e. the one from which the vector of sample proportions is computed). Bross gives a useful characterization of the mean ridit of a new group in this context: it can be viewed as the probability that, randomly sampling an individual from the new group, he presents a category in the ordered scale lower than the category presented by an individual randomly sampled from the reference group. Bross also constructed confidence intervals for the mean ridits of different classes of driver to draw inference about accident incidence. A simple way to use ridits in order to perform classical statistical analysis for comparing two groups of observations would be to determine the vector of sample proportions from the groups combined and then perform the two sample t-test after applying the scores to the groups separately. This can also be easily extended to the comparison of multiple groups.

Ridit have also a strict link with another scoring system, the mid-ranks. Mid-ranks are the averages of the ranks that would be assigned if the observations in a category could be ranked without ties. So for example the mid-rank for the first category m_1 is the average of the ranks $1, \ldots, n_1$ for the first n_1 respondents, so $m_1 = (1 + n_1)/2$. Whereas ridit scores fall between 0 and 1, mid-rank scores fall between 1 and N. The mid-rank for the i^{th} category is defined as:

$$m_i = \frac{\left[\left(\sum_{h=1}^{i-1} n_h\right) + 1\right] + \sum_{h=1}^{i} n_h}{2}$$

If the response probabilities are not theoretical but estimated from a sample of N respondent then there is a linear relationship between ridits and mid-ranks: we can obtain the former scores from the latter by subtracting $1/2$ and dividing by N. Selvin (1977) has also demonstrated that the mean ridit of a new group of observations is a linear transformation of the sum of category mid-ranks for this group when it is compared to the reference group as in the Wilcoxon rank sum test with many ties. The relation between ridits and mid-rank is the following:

$$r_i = \frac{m_i - 0.5}{N} \qquad i = 1, \ldots, k$$

As the mid-ranks are just a linear transformation of ridits the relative distances among categories between the two set of scores are the same and thus the classification results obtained would be identical. For these reasons we choose to not make use of mid-ranks in the simulations.

Equally spaced scores such raw scores or rank methods such ridits or mid-ranks are commonly used for processing ordinal data. However, if the data are right or left skewed or if some categories have many more observations than the others, using these method can lead to poor results. It could be more appropriate in these circumstances to rely on methods which incorporate some prior knowledge or expectation about the data distribution.

2.2.3 Normal median scores

The normal median scores, introduced by Brockett (1981), are chosen so that they minimize the Kolmogorov-Smirnov distance between F (observed cumulative distribution function) and G (theoretical cumulative distribution function underlying the ordinal variable):

$$d(F,G) = \underset{x}{max}|F(x) - G(x)|$$

The distance is minimized when $G(s_i) = r_i$, with $i = 1, \ldots, k$. Where r_i is the ridit score for category i.
So the scoring system is defined as:

$$s_i = G^{-1}(r_i) \quad i = 1, \ldots, k$$

We assume here a standard normal distribution underlying the observed ordinal scale. Thus, normal median scoring select s_i to minimize the distance to normality. If we take G as the uniform the scoring system become the Bross ridits.
In the normal median scores G is chosen to be the standard normal cumulative distribution function:

$$s_i = \Phi^{-1}(r_i) \quad i = 1, \ldots, k$$

The i^{th} score thus correspond to the r_i-centile of the standard normal distribution. In dealing with numerical values assigned to the categories of ordinal variables in standard parametric analysis one may face with large standard errors associated with parameter estimates. For how they are constructed this kind of scores should enhance the robustness of the analysis using linear models or in general any kind of analysis for which normality is an assumption.
As well as the ridits, the normal mean scores can be used in order to perform finer statisical analyses where data are measured on the ordinal scale such as in Educational and Psychological Testing. In fact, ordinal test items such as Likert scales result in raw scores that are meaningless without purposeful statistical interpretation. Normal mean scores, as well as other kind of scores (see blom scores) allow to modify raw scores values mathematically through a first step of standardization based on ranks that enables statistical procedures and a second step of normalization, needed for meaningful comparisons between scales.

2.2.4 Blom scores

Blom scores are very close to normal median scores as they both try to approximate the percentiles of the standard normal distribution.
The scores proposed by Blom originate from the problem of order statistic discussed originally by Pearson (1902), who provides a solution of a problem proposed by Galton (1902) of finding the average difference between two individuals in a ordered sample of size N. As the knowledge of average differences for symmetric populations also involves knowledge of the expected values of all the order statistic, several authors provide tables of such expected values, more or less accurate (see Harter, 1961 for a summary of these works). The Blom's proposed method for approximating the i^{th} smallest value normal order statistic for a sample of size N is the following:

$$s_i = \Phi^{-1}\left(\frac{m_i - \alpha}{N - 2\alpha + 1}\right) \qquad i = 1, \ldots, k$$

where m_i is the mid-rank value for category i.
Blom's formula responds to the curvilinear relationship between a score's rank in a sample and its normal variable.
In order to find the best value for α Blom tabulated the values required to yield the correct expectation of the i^{th} order statistic for i going from 1 to $N/2$ (because we deal with symmetric population the knowledge of the first $N/2$ expected values its sufficient as we can obtain the other half just by switching the sign) and N going from 2 to 400. He noticed that α ranges between 0.330 (for $N = 2$ and $i = 1$) and 0.5. In particular α increases as N increases and, for fixed N, α is minimum for $i = 1$, then rises quickly to a peak before dropping off slowly as i increases too much.
For this reason Blom suggested to use $\alpha = 3/8$ as a compromising value, yielding the scores:

$$s_i = \Phi^{-1}\left(\frac{m_i - 0.375}{N + 0.25}\right) \qquad i = 1, \ldots, k$$

Several rank-based normalization procedures have been developed among the years from ordinal data in order to perform analyses (see among the others the Van der Waerden, Rankit and Tukey procedures, all cited in Solomon & Sawilowsky, 2009). Arithmetically, all these methods do not differ substantially from blom scores, which are our choice in this study. We do not expect significant differences in the classifiers performance going from one procedure to the other as numerical values assigned to the categories do not differ if not for decimals. We leave to further studies the inclusion of these methods for comparative purposes.

2.2.5 Conditional mean scores

Conditional mean scores have been introduced by Brockett (1981) and then, independently and in a different formulation, by Fielding (1993). As in normal median and blom scores here it is assumed that the categories reflect an underlying continuous latent variable with cumulative distribution function G. The idea is to estimate the conditional mean of responses in the group under the assumed distributional form. The expression of the score, conditionally to the i^{th} category is:

$$s_i = \frac{1}{p_i} \int_{G^{-1}(\pi_{i-1})}^{G^{-1}(\pi_i)} x \, dG(x) = \frac{1}{p_i} \int_{\pi_{i-1}}^{\pi_i} G^{-1}(u) \, du \tag{2.1}$$

G denotes some given cumulative distribution, selected either in accordance with some theoretical latent distribution of the categorical variable under study or in accordance with the desirable properties for the planned method of analysis. Different choice of G lead to different sets of scores.
G^{-1} is the corresponding quantile function of the latent distribution, evaluated in correspondence of the cumulative sample proportions π_i for $i = 2, \ldots, k$.
Fielding (1997) developed an axiomatic framework for ensuring that the conditional mean scoring functions satisfy a set of postulates that a reasonable scoring method should posses. These are as follows:

Postulate 1. $h_1(1,1) = 0$; that is, in the case of one category the score can arbitrary set at 0.

Postulate 2. $0 \leq h_2(2, p, 1-p) = -h_2(1, 1-p, p)$; this reflects the idea that in the case of two categories if the distribution is reversed than by symmetry the absolute numerical values of the scores should be switched.

Postulate 3. If we are in the case of $k > 2$ and two adjacent categories are combined then the remaining scores should stay unchanged. That is, if the i^{th} and $(i+1)^{th}$ categories are combined into one then:

$$h_{k-1}(t, p_1, \ldots, p_{i-1}, p_i{+}p_{i+1}, p_{i+2}, \ldots, p_k) = \begin{cases} h_k(t, p_1, \ldots, p_k), & \text{if } t \leq i-1 \\ h_k(t+1, p_1, \ldots, p_k), & \text{if } t \geq i+1 \end{cases}$$

This postulate basically states that there is consistency in the scoring system as k changes.

Postulate 4. h_k is a bounded continuous function of the elements $(p_1, \ldots, p_k)$. This ensures that small changes in the sample proportions do not change the scores too much.

Postulate 5. If we are in the case of $k > 2$ and two adjacent categories are combined then the overall mean score is unaffected.

Postulate 6. $h_1(1, p_1, \ldots, p_k) \leq h_2(2, p_1, \ldots, p_k) \leq \ldots \leq h_k(k, p_1, \ldots, p_k)$; this means that the scores should reflect the order among the categories.

Fielding (1997) has carried on the previous work of Brockett & Levine (1977), where is shown that postulates 1-4, together with further recruitment that for $k = 2$ the expression $h_2(2, p, 1-p) - h_2(1, 1-p, p)$ is non-decreasing in p, lead to ridit scores. Fielding has shown that this was erroneous and that the latter postulate was unnecessarily restrictive for a reasonable scoring system. The ridits can be seen as a special case of conditional mean scores when the underlying distribution is assumed to be uniform.
Although these are undoubtedly appropriate postulates for some kind of statistical analysis, we will not consider in this study any situation where they can be useful with the only exceptions of postulates 4 and 6.

We consider in the present work three probability density functions for the underlying latent variable, among the most used in the literature and which seem to us useful in many practical problems, namely: the standard normal, the logistic and the log-normal. In considering an underlying skewed distribution such as the log-normal distribution we are violating the postulate 2. This postulate may not be necessary in this case. Indeed, it appears reasonable to consider an underlying distribution for possibly skewed observed data. Fielding (1997) has suggested that what motivate the introduction of postulate 2 (in Brockett & Levine, 1977) was the implicit assumption that the underlying variable has a symmetrical distribution but this is not desirable in any circumstance. He demonstrated that taking G as the log-normal cumulative distribution function the other postulates, and in particular the order condition among the scores, were still fulfilled by the scoring system.
As in the classification task what can possibly influence the classificator is the information about the relative distances within the categories, it appears

reasonable to think that any set of scores obtained by shifting the underlying distribution on the axis of abscissas would lead to the same result in terms of misclassification error. So, for reason of simplicity, we proceed considering underlying distributions centred at zero for scores based on normal and logistic latent distributions (i.e. the expected value $E(x) = \mu = 0$, where x is the respective underlying variable). For the scores based on the log-normal latent distribution we consider the logarithm of the underlying random variable, distributed according with a standard normal (thus with $E(\log(x)) = 0$, where x is log-normally distributed).
The scores for the i^{th} category are computed as following:

1. Normal mean scores (NMS) arises when G is the cumulative distribution function of the standard normal distribution. Following the notation in Brockett (1981), we first define $Z(\pi) = \Phi^{-1}(\pi)$ to be the Probit function. The scores are then:

$$\begin{aligned} s_i &= \frac{1}{p_i}\int_{\pi_{i-1}}^{\pi_i} Z(u)\,du \\ &= \frac{1}{p_i}\int_{\pi_{i-1}}^{\pi_i} \sqrt{2}\texttt{erf}^{-1}(u)\,du \\ &= \frac{1}{p_i}\left\{\phi\left[Z(\pi_{i-1})\right] - \phi\left[Z(\pi_i)\right]\right\} \\ &= \frac{\exp\left(-Z^2(\pi_{i-1})/2\right) - \exp\left(-Z^2(\pi_i)/2\right)}{p_i\sqrt{2\pi}} \end{aligned}$$

   for $i = 2, \ldots, k$.
   Where $\texttt{erf}(x) = (2/\sqrt{\pi})\int_0^x e^{-t^2}\partial t$ is the Gauss error function and ϕ is the standard normal probability density function. For $i = 1$ the score is:

$$s_1 = \frac{-\exp\left(-Z^2(\pi_1)/2\right)}{p_1\sqrt{2\pi}}$$

2. The logistic mean scores (LMS) are computed setting $G(x) = 1/(1 + e^{-x})$ (cumulative distribution function of the logistic distribution with

mean 0 and variance $\pi^2/3$), this yields scores:

$$\begin{aligned}
s_i &= \frac{1}{p_i}\int_{\pi_{i-1}}^{\pi_i} \log\left(\frac{u}{1-u}\right) du \\
&= \frac{1}{p_i}\left[\pi_i \log\left(\frac{\pi_i}{1-\pi_i}\right) - \pi_{i-1}\log\left(\frac{\pi_{i-1}}{1-\pi_{i-1}}\right) + \log\left(\frac{\pi_{i-1}}{1-\pi_{i-1}}\right)\right] \\
&= \frac{H(\pi_i) - H(\pi_{i-1})}{p_i}
\end{aligned}$$

where $H(\pi) = \pi\log(\pi) + (1-\pi)\log(\pi)$ is the entropy function. For $i = 1$ and $i = k$ the scores are, respectively:

$$s_1 = \frac{H(\pi_1)}{p_1} \qquad\qquad s_k = \frac{-H(\pi_{k-1})}{p_k}$$

3. The log-normal mean scores (LNMS) are computed setting $G(x) = \Phi(\log(x)) = \frac{1}{2}\left(\frac{1}{2} + \texttt{erf}\left[\frac{\log(x)}{\sqrt{2}}\right]\right)$, cumulative distribution function of $x \in (0, \infty)$, where $\log(x)$ is the standard normal distribution. The scores are as follows:

$$\begin{aligned}
s_i &= \frac{1}{p_i}\int_{\pi_{i-1}}^{\pi_i} \exp\left(\Phi^{-1}(u)\right) du \\
&= \frac{\Phi\left[\Phi^{-1}(\pi_i) - 1\right] - \Phi\left[\Phi^{-1}(\pi_{i-1}) - 1\right]}{p_i}
\end{aligned}$$

For $i = 1$ the score is:

$$s_1 = \frac{\Phi\left[\Phi^{-1}(\pi_1) - 1\right]}{p_1}$$

Since these scores, if not based upon theoretical probabilities associated with the variable categories, are functions of the observed sample proportions, it appears to us that an appropriate choice of the score could not be clear in any circumstances. This choice may depend on the data set upon which classification is based. Therefore, we proceed by performing a sensitivity analysis, that is choose scores in the different ways just presented, that are not linear transformations, and check whether conclusions depends on the chosen set of scores.

As one may notice the most part of the implemented scoring methods used in this study assume a normal or uniform latent distribution. This is not an unreasonable assumption and in many cases may lead to fruitful results. A further justification for this is that each variable in the dataset can be seen as a mixture of observations from different populations thus, if the number of categories is low (generally for ordinal variables the classical examples include 5, 7 or 9 categories) it is possible that, by analysing the observations belonging to each class separately, these show specific (and possibly skewed) distributions but considering the distribution of all the observations jointly a sort of "masking" effect may happen that can lead to prefer scores that do not assume any particular information from the data. However, small samples and skewed distributions are likely to degrade the performance of all methods so we include also the case of log-normal latent distribution.

Before going further we would like to point out that all the scoring methods have undoubtedly advantages because, as mentioned, they allow a direct application of classical statistical analyses, allowing to overcome, with a choice lead-by-data, the problems related to data collected on an ordinal scale. In addition, these methods make comparisons between groups and descriptive statistics possible. Particularly with regard to ridits, mid-ranks and normal median scores, their simplicity of interpretation and of use make them a suitable tool also for non-technical researchers.
Since the application of the presented scoring methods is not routine in available software, dedicated algorithms have been developed using the statistical environment **R**. With regard to the computational times, we present in Table 2.3 the times (in seconds) required to the statistical software **R** for computing the scores. The table shows the elapsed time in **R** for running each scoring method 100 times. The ordinal data upon which scoring methods have been applied has been obtained by categorizing a continuous vector randomly sampled from a standard normal variable. Computational times are reported for different dataset sizes (N) and number of categories (C).
As it is possible to notice in every case the time required for computing scores is extremely low, which is a clear advantage. Times required for computing normal median scores and blom scores are generally higher than, respectively, ridits and mid-ranks as they involve the computation of these last two kind of scores (i.e. in order to obtain normal median scores, ridits have to be computed and, similarly, to obtain blom scores, mid-ranks have to be computed). Furthermore, as one would expect, computational time increases as the dimension of the data set and/or the number of categories increases.

Table 2.3: *Computational times (in seconds) required for computing each scoring method 100 times. The results are displayed for different numbers of observations (N) and numbers of categories of the ordinal variable (C).*

	N=100		
	C=5	**C=7**	**C=9**
Ridits	0.14	0.17	0.19
Midranks	0.32	0.42	0.56
Normal Median Scores	0.15	0.17	0.20
Blom Scores	0.33	0.44	0.56
Normal Mean Scores	0.14	0.17	0.20
Logistic Mean Scores	0.12	0.16	0.18
Log-Normal Mean Scores	0.14	0.17	0.22
	N=200		
	C=5	**C=7**	**C=9**
Ridits	0.23	0.27	0.32
Midranks	0.50	0.64	0.85
Normal Median Scores	0.23	0.28	0.33
Blom Scores	0.49	0.67	0.86
Normal Mean Scores	0.22	0.27	0.33
Logistic Mean Scores	0.22	0.25	0.31
Log-Normal Mean Scores	0.20	0.26	0.32
	N=400		
	C=5	**C=7**	**C=9**
Ridits	0.42	0.48	0.57
Midranks	0.87	1.14	1.41
Normal Median Scores	0.42	0.52	0.59
Blom Scores	0.85	1.14	1.42
Normal Mean Scores	0.39	0.47	0.55
Logistic Mean Scores	0.39	0.47	0.55
Log-Normal Mean Scores	0.40	0.47	0.56

To better understand the differences between score types we also present in Figure 2.1 a graphical comparison of raw scores (on the abscissa axis) with the other methods of scoring. The reference ordinal variable has been generated, as for the computational times, randomly sampling from a standard normal variable, which has been subsequently discretized, considering five categories (more details on the method used in order to generate ordinal variables will be provided in chapter 5). The scores have been computed on a sample of 200 observations.

We do not include in the graphs the comparison with mid-ranks as they are just a linear transformation of ridits, so the relative distances between categories would be the same. Mid-ranks are included in the Table 2.4, where the sample proportions, together with the numeric values for the scores presented in the graphs are provided.

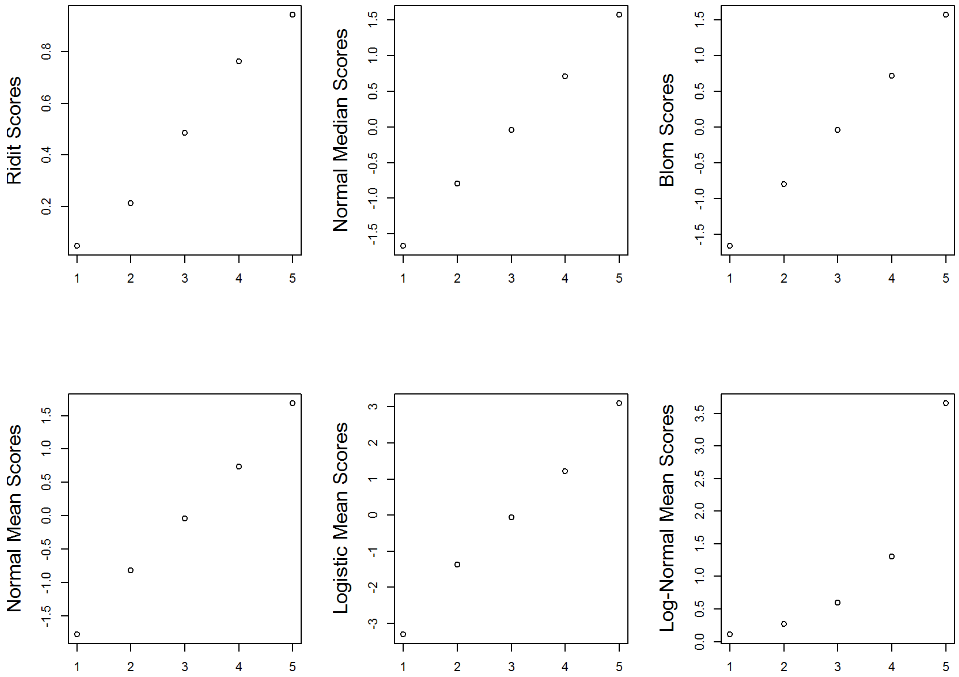

Figure 2.1: *Comparison of scoring methods with raw scores.*

As expected, we do not observe relevant differences between normal median and blom scores from the graph in Figure 2.1 (differences are at the third decimal place as it possible to see from Table 2.4). With regard to normal mean and logistic mean scores we have that the scores for the high order categories are more spread for the latter method. The exception is category 3 (i.e. the central category), where the scores are similar due to the fact that both the underlying distributions are centred at 0. This is because the logistic distribution has slightly longer tails compared to the normal distribution. The main difference in the score's trend is observed for the log-normal mean scores. In this case the scores are closer for the low order categories and the distance increases for high order categories. When data are generated as in this example, by discretizing from a continuous standard normal variable, we may expect that, in a classification context, using log-normal mean scores would bring the worst result with respect to all other scores.

Table 2.4: *Score's value for simulated standard normal data*

	Category 1	**Category 2**	**Category 3**	**Category 4**	**Category 5**
Frequencies	19	47	62	49	23
Sample Proportions	0.095	0.235	0.310	0.245	0.115
Raw Scores	1	2	3	4	5
Ridit Scores	0.0475	0.2125	0.4850	0.7625	0.9425
Mid-Rank Scores	10	43	97.5	153	189
Normal Median Scores	-1.6696	-0.7978	-0.0376	0.7144	1.5761
Blom Scores	-1.6639	-0.7965	-0.0376	0.7133	1.5713
Normal Mean Scores	-1.7791	-0.8218	-0.0386	0.7348	1.6878
Logistic Mean Scores	-3.3048	-1.3627	-0.0621	1.2105	3.1030
Log-Normal Mean Scores	0.1098	0.2745	0.5988	1.3013	3.6574

Again we remark that the location of the underlying curve is not of interest in this context as long as the relative differences between categories stay the same. A justification for this come from the simple realization that any classifier should be able to identify the best subdivision region among observations into classes, regardless of translations of the data values (i.e. a reasonable classifier should be invariant with respect to linear transformations). For example, suppose that we are in the situation of univariate classification (with two classes) with data coded as ridits scores in Table 2.4. If, from the chosen classifier, it results that the optimal misclassification rate is obtained by assigning "class 1" to observations that fall in category 1 or 2 (0.0475 or 0.2125 from the table) and assigning "class 2" to the rest, then if we move to another set of scores by adding to each category a constant value k, with $k \in \mathbb{R}$, the optimal misclassification rate should come as result of the same subdivision as before, whatever the value of k is.

As mentioned, all the scores in the table refer to the application to a single ordinal variable. If the number of variables on which we rely is higher, it is not generally possible to apply the presented methods on the whole set of ordinal variables (for example, it may happen that the variables have not the same number of categories). The scores have to be applied separately to each variable.

For this reason, a strong limitation that emerge in using these scores is that there is a loss of information due to the fact that the dependence structure between variables is ignored. Correlated variables are not considered in the scores computation but they may have a significant effect on the classifier's performance. Moreover, as it is possible to see from Table 2.4, different distributional forms hypothesized for the underlying variable lead to different score values, thus the relative distances among two pairs of adjacent categories also change. Therefore, if the underlying distribution is misspecified there is a possibility of high classification errors.

In the next chapter we will present our proposal, based on the response function approach, which has the aim of overcome the issues associated with the scoring methods.

Chapter 3

Our Proposal

3.1 The idea

Despite the fact that scoring methods presented in chapter 2 are definitely a useful tool as they allow to treat ordinal data directly in a classification framework, they also present serious limitations as:

1. They do not allow to treat all variables simultaneously, thus there is a loss of information due to possible correlations among variables. The assumption of correlated observed variables in a classification context is not unreasonable. If the features we observe are informative for classification, we may expect that they are correlated to each other.

2. If we have knowledge about the distribution of the population from which the data come, we may hypothesize a suitable functional form for the distributions underlying each variable in the data set upon which scores are computed, with the possibility of having good classification results. In many cases, however, we do not have such knowledge and we have to rely on the empirical observations, which may be misleading. In these cases, it would be preferable to consider an underlying distribution in order to minimize the classification error, assuming that there is no information from the data.

In order to overcome these issues, we propose to use the response function approach introduced by Moustaki (2000), developed within the framework of Generalized Linear Latent Variable Models (GLLVM). As shown later, using the response function approach allows to move from the problem of classifying

a set of ordinal variables to classifying a set of continuous variables, smaller in number, where traditional methods of classification can be applied.
Cagnone & Viroli (2014) presented an innovative method where the response function approach is applied considering as latent distribution a mixture of multivariate Gaussians. This allows for a more flexible latent distribution that may better fit the data in classification tasks where observations come from heterogeneous sub-populations. However, this method also presents some serious limitations due to the large number of parameters to estimate.
The main idea of the present work is to avoid the limitations that emerge in using the mixture of multivariate Gaussian distributions, considering for the classification task a Beta latent distribution. It will be shown that it allows not only for more flexibility but also for faster computational times.

Latent Variable Models

Latent Variable Models (LVM) generally arise to explain the interdependence within a large set of variables through a small number of underlying non-observable factors, uncorrelated each other in the sense that, if the factors were fixed and known, the observed (manifest) variables would be independent. The basic idea is to summarize the information contained in a given set of response variables $x_1, \dots, x_p$ with a set of latent factors $z_1, \dots, z_q$, usually assumed to be much smaller in number than the observable variables.
Dependently on the nature of the latent and manifest variables, it is possible to distinguish among four different kind of analyses, as shown in Table 3.1. When both latent and manifest variables are continuous the latent model is the well-known classical factor analysis. When the latent variables are discrete then we have the latent profile analysis (for continuous manifest variables) or latent class analysis (for discrete manifest variables). In our case the aim is to move the classification problem towards an interval measurement scale so that standard classification methods can be applied, thus we are performing a latent trait analysis as we consider continuous latent variables and discrete manifest variables.

Table 3.1: *Latent variable models classification.*

	Latent Variables	
Manifest variables	Continuous	Discrete
Continuous	Factor Analysis	Latent Profile Analysis
Discrete	Latent Trait Analysis	Latent Class Analysis

In dealing with ordinal variables using latent variable models we can distinguish between two main approaches: The Underlying Response Variable Approach (URVA), which is the most popular one (Jöreskog, 1990 and Muthén, 1984) and will be described briefly below, and the response function approach, which is the method we choose and it will be described more in detail in the next section. An overview of those type of models can be found in Knott & Bartholomew (1999).

Before introducing the URVA and response function approach we introduce the general framework upon which we base the computations.

General Framework

Let:

- $\mathbf{x} = (x_1, \ldots, x_p)$ be the set of p observed ordinal variables;
- k_i number of categories for the i^{th} observed variable, $(i = 1, \ldots, p)$;
- $x_{hi} = a_i$ be the h^{th} observation of the i^{th} ordinal variable belonging to the ordered category a, for $a_i = 1, \ldots, k_i$. Thus, $\mathbf{x}_h = (x_{h1} = a_1, x_{h2} = a_2, \ldots, x_{hp} = a_p)$ is the response pattern associated with the h^{th} unit $(h = 1, \ldots, N)$. There are then $\prod_1^p k_i$ possible response patterns;
- $\mathbf{z} = (z_1, \ldots, z_q)$ be the set of q latent variables, with $q < p$.

The underlying response variable approach (URVA)

The URVA follows the classical factor analysis model, assuming that each ordinal variable comes as a discretization of an underlying Gaussian latent variable. The origin of factor analysis is generally ascribed to Charles Spearman, who first developed a general framework in the psychometric field with a single general factor and a number of specific factors (Spearman, 1904). The model has the form:

$$x_i^* = \alpha_i + \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \ldots + \lambda_{iq}\xi_q + u_i \quad i = 1, \ldots, p \tag{3.1}$$

where:

- α_i is the mean value of x_i^*, for $i = 1, \ldots, p$.
- λ_{ij}, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, are the factor loadings as they express the load of each factor on the observed variable.
- $\xi_1, \ldots, \xi_q$ are the latent variables or common factors. In the standard factor model it is assumed that the factors have zero mean, unitary variance (because they do not have a given unit of measure and scale) and that they are uncorrelated.
- u_i are the error terms representing a specific factor and measurement error. They are called unique factors and are assumed to be independent and normally distributed as $u_i \sim N(0, \psi_i)$ for $i = 1, \ldots, p$.

Equation (3.1) is a suitable representation of a factor analysis model if the observed variables are continuous (Table 3.1). This model is not appropriate when the observed variables are ordinal. For this reason, in the model x_i^* is considered to be a standard normal variable underlying the ordinal variable x_i. We define x_i^* such that:

$$x_i = \begin{cases} 1 & if \;\; x_i^* \geq \tau_{i1} \\ 2 & if \;\; \tau_{i1} < x_i^* \leq \tau_{i2} \\ \vdots & \\ k_i & if \;\; \tau_{ik_i-1} < x_i^* \end{cases}$$

The parameters $\tau_{i1}, \tau_{i2}, \ldots, \tau_{ik_i-1}$ are called threshold values. For each observed variable x_i with k_i categories there are $k_i - 1$ related thresholds. The thresholds reflect the order condition of the categories:

$$\tau_{i0} = -\infty, \tau_{i1} < \tau_{i2} < \ldots < \tau_{ik_i-1}, \tau_{ik_i} = +\infty$$

thus, the probability of a general p-dimensional response pattern $\mathbf{x}_h = (x_{h1} = a_1, x_{h2} = a_2, \ldots, x_{hp} = a_p)$ is given by:

$$P(\mathbf{x}_h) = \int_{\tau_{1a_1-1}}^{\tau_{1a_1}} \int_{\tau_{2a_2-1}}^{\tau_{2a_2}} \cdots \int_{\tau_{pa_p-1}}^{\tau_{pa_p}} \phi_p\left(t_1, t_2 \ldots, t_p | \boldsymbol{P}\right) \, dt_1, dt_2, \ldots, dt_p, \tag{3.2}$$

where ϕ_p is a p-dimensional normal density function with zero means, unit variances and correlation matrix $\boldsymbol{P}$.
The URVA is a full information maximum likelihood approach applied to the factor analysis model. Parameters estimation requires the evaluation of a p-dimensional integral (equation (3.2)). Computational times increase rapidly with p so this approach is not computationally feasible. Other methods have been proposed such as a limited information method for estimating the model parameters, which maximizes the sum of all univariate and bivariate marginal likelihoods (Jöreskog & Moustaki, 2001).
Several authors proposed alternative methods for estimating the model parameters. These methods differentiate for the number of stages required to obtain the estimates, usually two or three. Examples of three stage methods are:

- Muthén (1984):
    1. Estimation of first order statistic (means, variances, thresholds, etc.) by maximum likelihood.
    2. Estimation of second order statistics such as polychoric correlations by conditional maximum likelihood.
    3. Parameters of the stuctural part of the model are estimated using a limited information generalized least squares method.
- Jöreskog (1994)
    1. Estimation of first order statistic (means, variances, thresholds, etc.) by maximum likelihood.
    2. Estimation of second order statistics such as polychoric correlations by conditional maximum likelihood.
    3. Parameters of the stuctural part of the model are estimated using weighted least squares (the weight matrix correspond to the inverse of asymptotic covariance matrix of polychoric correlations).

Example of two stage estimation:

- Lee *et al.* (1995):
    1. First stage estimates like thresholds, polychoric and polyserial correlations in the underlying correlation matrix are obtained using a partition maximum likelihood approach.

2. A generalized least squares approach is employed to estimate the structural parameters in the correlation structure on the basis of the joint asymptotic distribution of the first stage estimator and a weight matrix.

3.2 The response function approach

The response function approach is particularly used within item response theory with binary variables and a single latent factor (Jöreskog & Moustaki, 2001). Contrary to the URVA, an underlying variable is not defined for each observed ordinal variable, instead the unit of analysis is the complete p-dimensional response pattern distribution, conditionally on the latent factors, so there is no loss of information due to correlations within features. It is assumed that responses to different variables are independent for given latent factors.

This approach is a generalization in the context of latent variable models of the class of models discussed by McCullagh (1980), which include proportional odds model and probit model. These models are in turn a multivariate extension of generalized linear models.

Basing on the work of Moustaki (2003) and Moustaki & Knott (2000), we considered a generalized latent variable model for ordinal data.

3.2.1 Measurement model

The joint density function of the ordinal variables $\mathbf{x}$ is given by:

$$f(\mathbf{x}) = \int_{\mathrm{I\!R}^q} f(\mathbf{x}, \mathbf{z})\, dz = \int_{\mathrm{I\!R}^q} f(\mathbf{x}|\mathbf{z}) f(\mathbf{z})\, dz, \tag{3.3}$$

where $\mathbf{z}$ is the set of q latent variables.

Under the assumption of conditional independence, $\mathbf{z}$ accounts for the association among ordinal variables. Conditional independence is a necessary condition so that the conditioning variables provide an adequate explanation of the correlation between the ordinal variables. For fixed $\mathbf{z}$ we have:

$$f(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{p} f_i(x_i|\mathbf{z}).$$

A generalized linear model consist of three components (McCullagh, 1984):

1. a random component with distribution that belongs to the exponential family. It is generally assumed that each of the p manifest variables has a distribution of the form:

$$f_i(x_i;\theta_i;\phi_i) = \exp\left[\frac{x_i\theta_i - b_i(\theta_i)}{\phi_i} + c_i(x_i;\phi_i)\right], \quad i = 1,\ldots,p \tag{3.4}$$

   where θ_i is the canonical parameter, ϕ_i is the scale parameter, $b_i(\theta_i)$ and $c_i(x_i;\phi_i)$ are specific functions, the former depends only on the parameters while the latter depends on the parameters and the observed data;

2. a systematic component, in which the latent variables produce a linear predictor for each observed variable x_i:

$$\eta_i = \lambda_{i0} - \sum_{j=1}^{q}\lambda_{ij}z_j, \quad i = 1,\ldots,p;$$

3. the link between the systematic component and the conditional mean of the random component:

$$\eta_i = g_i(\mu_i(\mathbf{z})), \quad i = 1,\ldots,p$$

   where $\mu_i(\mathbf{z}) = E(x_i|\mathbf{z})$ and g_i is the link function. The link function can be any monotone and differentiable function, whose domain is in the range $[0,1]$ and that assumes values in $[-\infty,\infty]$. Examples of link functions are the logit, the probit, the complementary log-log function and the log-log function.

Since we are dealing with ordinal data the random component in the model corresponds to the multinomial distribution. We define:

- $\pi_{i(l)}(\mathbf{z}) = P(x_i = l|\mathbf{z})$ the conditional probability for the i^{th} ordinal variable to assume category l.

- $\gamma_{i(l)}(\mathbf{z}) = P(x_i \leq l|\mathbf{z}) = \pi_{i(1)}(\mathbf{z}) + \ldots + \pi_{i(l)}(\mathbf{z})$ the cumulative probability of a response to be in category l or lower for the i^{th} observed variable.

The conditional probability of the i^{th} observed variable is then:

$$f_i(x_i|\mathbf{z}) = \prod_{l=1}^{k_i}\pi_{i(l)}(\mathbf{z})^{x_{i(l)}} = \prod_{l=1}^{k_i}\left[\gamma_{i(l)}(\mathbf{z}) - \gamma_{i(l-1)}(\mathbf{z})\right]^{x_{i(l)}} \tag{3.5}$$

where $x_{i(l)} = 1$ if an observation is in category l for the i^{th} observed variable and $x_{i(l)} = 0$ otherwise. Equation (3.5) can be rephrased as (we denote for simplicity in notation $\gamma_{i(l)}(\mathbf{z}) = \gamma_{i(l)}$):

$$f_i(x_i|\mathbf{z}) = \prod_{l=1}^{k_i-1} \left(\frac{\gamma_{i(l)}}{\gamma_{i(l+1)}}\right)^{x_{i(l)}} \left(\frac{\gamma_{i(l+1)} - \gamma_{i(l)}}{\gamma_{i(l+1)}}\right)^{x_{i(l+1)}-x_{i(l)}}. \tag{3.6}$$

If we apply the logarithm to equation (3.6) we have:

$$\log f_i(x_i|\mathbf{z}) = \sum_{l=1}^{k_i-1} \left(x_{i(l)} \log \frac{\gamma_{i(l)}}{\gamma_{i(l+1)} - \gamma_{i(l)}} - x_{i(l+1)} \log \frac{\gamma_{i(l+1)}}{\gamma_{i(l+1)} - \gamma_{i(l)}}\right). \tag{3.7}$$

Then the conditional probability for the i^{th} observed variable can be expressed in the general form of the exponential family distribution of equation (3.4):

$$f_i(x_i|\mathbf{z}) = \exp\left[\sum_{l=1}^{k_i-1} \left(x_{i(l)} \log \frac{\gamma_{i(l)}}{\gamma_{i(l+1)} - \gamma_{i(l)}} - x_{i(l+1)} \log \frac{\gamma_{i(l+1)}}{\gamma_{i(l+1)} - \gamma_{i(l)}}\right)\right] \tag{3.8}$$

where:

$$\theta_{i(l)}(\mathbf{z}) = \log \frac{\gamma_{i(l)}}{\gamma_{i(l+1)} - \gamma_{i(l)}}, \qquad l = 1, \ldots, k_i - 1,$$

and

$$b_i[\theta_{i(l)}(\mathbf{z})] = \log \frac{\gamma_{i(l+1)}}{\gamma_{i(l+1)} - \gamma_{i(l)}} = \log\{1 + \exp\left[\theta_{i(l)}(\mathbf{z})\right]\}, \qquad l = 1, \ldots, k_i - 1.$$

The systematic component of the model is category-dependent and is of the form:

$$\eta_{i(l)} = link\left[\gamma_{i(l)}(\mathbf{z})\right] = \lambda_{i0(l)} - \sum_{j=1}^{q} \lambda_{ij} z_j,$$

where $\lambda_{i(l)}$ are category-specific intercepts (or thresholds) with $\lambda_{i(1)} < \ldots < \lambda_{i(k_i)} = +\infty$ and the λ_{ij}, for $i = 1, \ldots, p$ and $j = 1, \ldots, q$, can be considered as factor loadings because they measure the effect of the latent variables $\mathbf{z}$ on the probability of responding in some category of the observed ordinal variable $\gamma_{i(l)}(\mathbf{z})$. The negative sign before λ_{ij} means that, if the loading has positive value, as z_j increases it is more likely for x_i to fall into higher category. The link function considered in this study is the logit:

$$logit\left[\gamma_{i(l)}(\mathbf{z})\right] = \lambda_{i0(l)} - \sum_{j=1}^{q} \lambda_{ij} z_j$$

thus, the cumulative probability can be expressed as:

$$\gamma_{i(l)}(\mathbf{z}) = \frac{\exp\left[\lambda_{i0(l)} - \sum_{j=1}^{q} \lambda_{ij} z_j\right]}{1 + \exp\left[\lambda_{i0(l)} - \sum_{j=1}^{q} \lambda_{ij} z_j\right]} = \Psi\left[\lambda_{i0(l)} - \sum_{j=1}^{q} \lambda_{ij} z_j\right]$$

where $\Psi(x)$ is the logistic distribution function.

Figure 3.1 shows an example where the cumulative probabilities $\gamma_{i(l)}$ for an ordinal variable with 5 categories are plotted against a single latent variable. The values of the latent variable z have been simulated sampling at random from a uniform distribution in the range $[-2, 2]$, while the values of the intercepts and factor loading have been generated sampling at random from a uniform distribution with domains $[-2, 2]$ and $[0, 3]$, respectively. The chosen values are reported in Table 3.2. Note that the threshold value for category $l = 5$ has not been simulated.

The size of the factor loading determines if the curves grow up or down and how quickly they do it. The curves for the cumulative probability have always the same order (corresponding to the order in the categories) whatever the values of z is.

Figure 3.2 shows the response probabilities $\pi_{i(l)}$ correspondent to the cumulative probabilities of Figure 3.1.

Table 3.2: *Simulated intercepts and factor loading.*

	Simulated values
$\lambda_{0(1)}$	-1.621337
$\lambda_{0(2)}$	-0.4437145
$\lambda_{0(3)}$	0.3322429
$\lambda_{0(4)}$	1.410525
$\lambda_{0(5)}$	$+\infty$
λ_1	2.36024

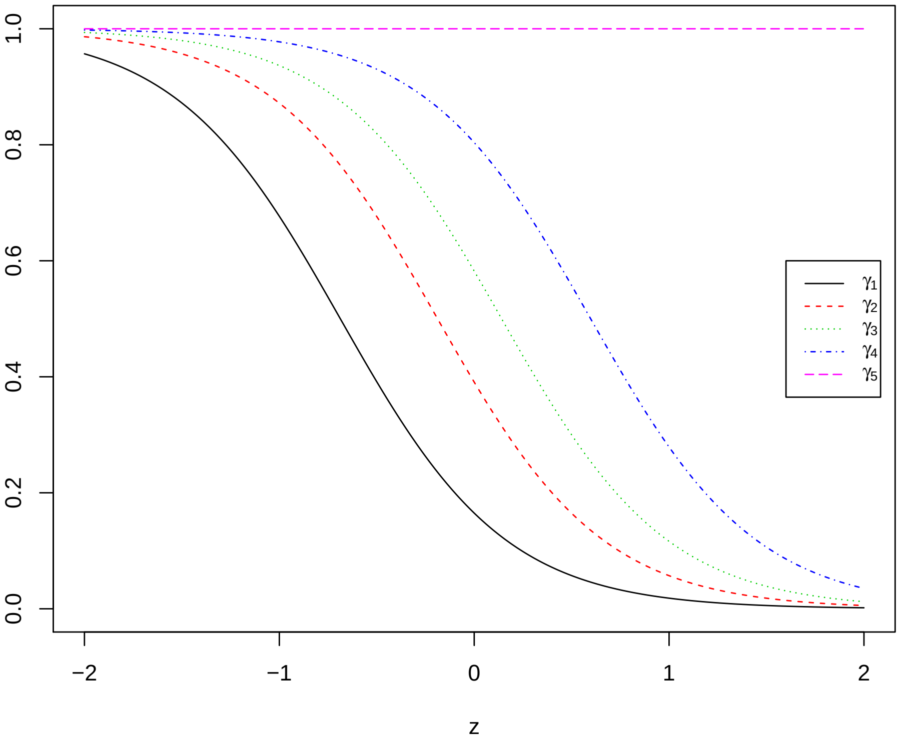

Figure 3.1: *Cumulative probabilities $\gamma_{(l)}$ (y-axis), $l = 1, \ldots, 5$, for different values of the latent variable z (x-axis). Cumulative probabilities have been obtained from the simulated values of thresholds and factor loading in Table 3.2.*

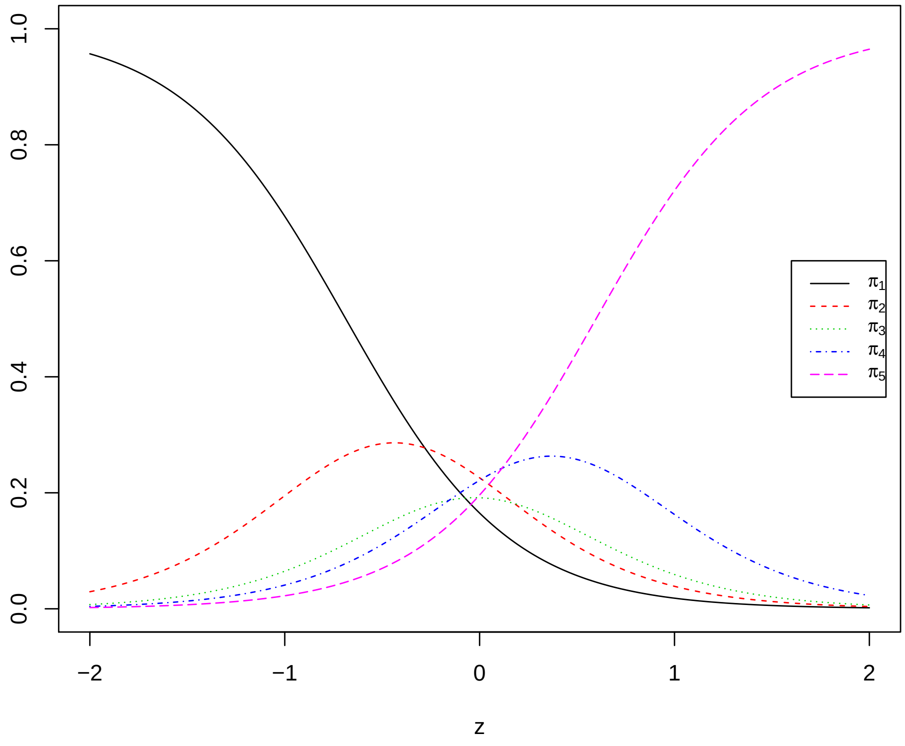

Figure 3.2: *Response probabilities $\pi_{i(l)}$ (y-axis), $l = 1, \ldots, 5$, for different values of the latent variable z (x-axis). The response probabilities correspond to the cumulative probabilities in Figure 3.1.*

3.2.2 Structural model

In the classical GLLVM $\mathbf{z}$ is assumed to be distributed according to a multivariate standard normal. Bartholomew (1988) suggested that the use of a standard normal distribution has rotational advantages when there is more than one latent variable.

However, the assumption of normally distributed latent variables cannot be appropriate for classification tasks where data come from an unobserved heterogeneous population.

Using alternatives to normality for the latent variables is not new in the statistical literature on GLLVM. See, among the others, Cagnone & Viroli (2014), Irincheeva *et al.* (2012), Wall *et al.* (2015), Montanari & Viroli (2010), Yung (1997) and Wedel & Kamakura (2001).

We present here an example to show the reason why a latent distribution that allows for more flexibility is more appealing when the task is classifying data coming from different sub-populations. As mentioned in chapter 1, we work within the underling variable approach, assuming a priori that the ordinal variables are the result of the discretization of an unmeasurable underlying variable. To exemplify this, Figure 3.3 shows the case where the underlying distribution is a χ^2 with 5 degrees of freedom for an ordinal variable with 5 categories. The threshold values have been chosen so that they are equidistant. In classification tasks, the data set is composed by observations which are the realization of a p-dimensional random variable with different and unknown class-dependent multivariate distributions. In our case then we can see the ordinal variables as coming from the discretization of a mixture of p-variate distributions.

Suppose that we are in the univariate case, with one manifest ordinal variable and one latent factor, where each observation can be assigned to one of three possible classes. The ordinal variable comes from the discretization of an underlying distribution, which is a mixture of three probability density functions corresponding to the three population densities.

Figure 3.4 shows an example where the underlying probability density functions associated to each populations are Gaussian distributions with means $\boldsymbol{\mu} = (0, 2, 5)$ and variances $\boldsymbol{\sigma^2} = (1, 1.5, 3)$, the bold black line represents the mixture density, supposing equal weights.

It seems reasonable then to assume that good results may be obtained only if the distribution of the latent variable, obtained through the response variable approach, to which the classifier is applied, approximates as well as possible the "true" underlying mixture distribution. In the example in Figure 3.4 it is possible to see that assuming the latent variable to be distributed according to a standard normal distribution is not the best possible solution as the mixture clearly shows asymmetry. Instead, it seems legit to choose a distributional form that allows for more flexibility.

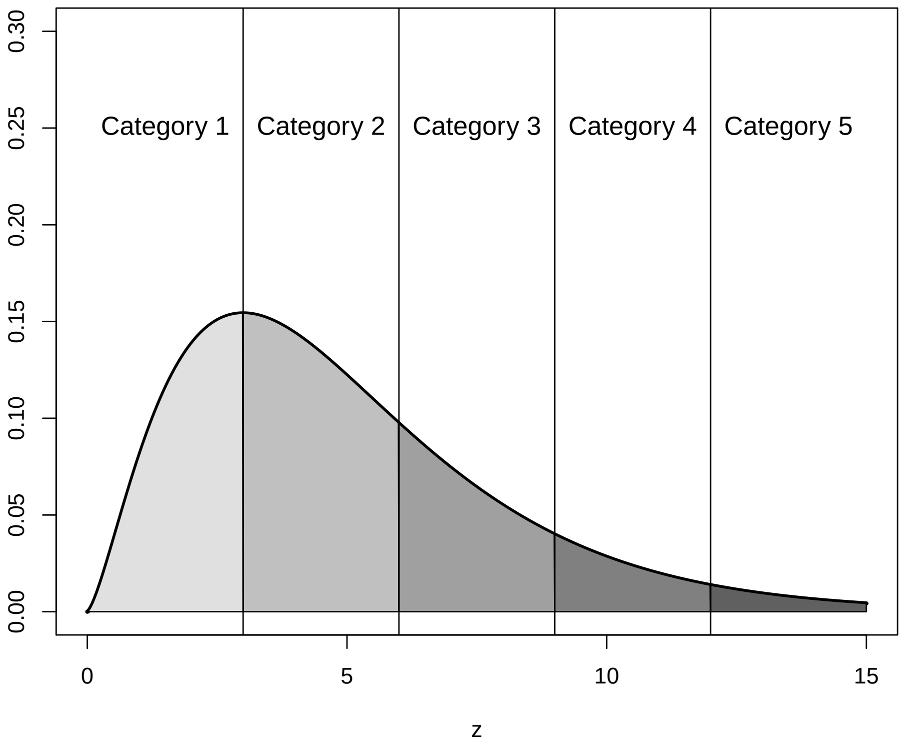

Figure 3.3: *Categories of an ordinal variable from a χ^2 underlying distribution with 5 degrees of freedom.*

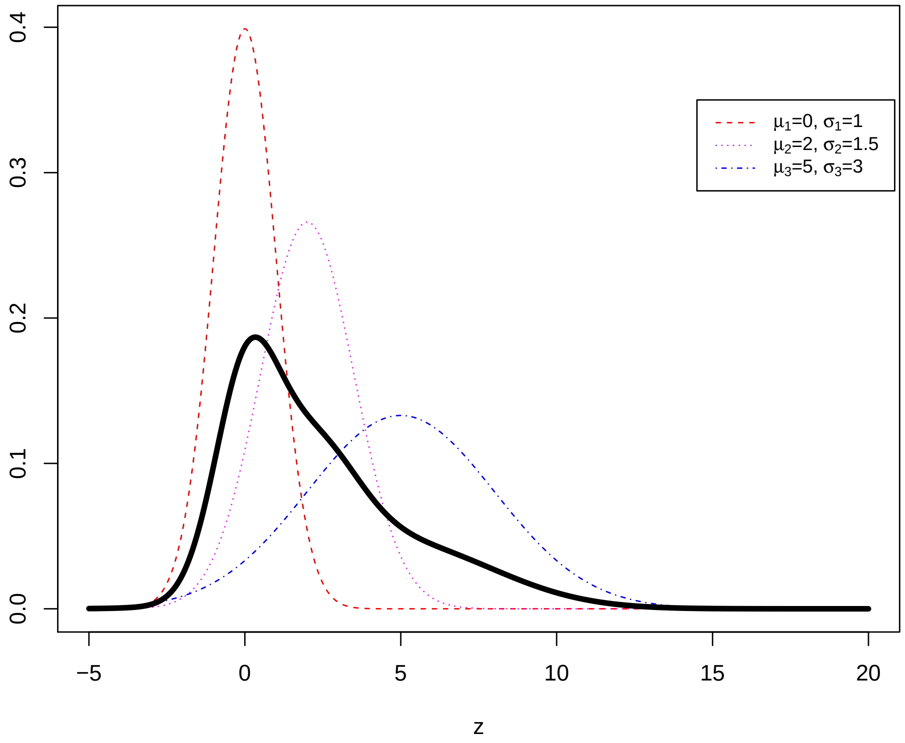

Figure 3.4: *Gaussian underlying probability density functions and associated mixture distribution (bold black line) with equal weights.*

Before going further in the description of the structural model we highlight that from now on we will consider only the case of a single latent variable. A motivation of this choice is related to the properties of a classifier that will be presented in chapter 4. Furthermore, this simplifies calculations and allows for faster computational times. We leave to future studies the development of the method for more than one latent factor.

3.3 A new proposal: The Beta Response Function Approach (BRFA)

The main idea proposed in this study is to introduce a Beta distributed latent variable in the context of the response function approach, presented in the previous section. As far as we know this distribution has never been applied in the latent variable models and, in particular, with the purpose of supervised classification of ordinal variables.

The idea of using a latent distribution that allows for more flexibility comes from the realization that the usual normality assumption of the latent variables is not always justified.

For this reason, in defining the structural part of the model we considered the work of Cagnone & Viroli (2014), where a mixture of multivariate Gaussians is proposed for the latent variables in order to account for heterogeneity in the data. This model comes as an extension of a previous work developed for binary data (Cagnone & Viroli, 2012). The aim of these models is twofold: they operate a dimension reduction in the input space and allow to perform a model based clustering in the latent space. Each cluster should correspond to one component of the mixture. With a single latent variable the structural part of the model takes the following form:

$$f(z) = \sum_{j=1}^{G} \pi_j \phi_j \left(\mu_j, \sigma_j\right) \qquad z \in [-\infty, +\infty]$$

where ϕ_j is a Gaussian density with mean μ_j and variance σ_j. The π_j are the mixing proportions which are unknown and have to be estimated.

The distributional form proposed by Cagnone and Viroli seems appealing in our case because, as mentioned, it allows to better approximate the case where data come as a mixture of class-conditional distributions.

Anyway, despite its simplicity and reasonably for application, this method presents some limitations. As the number of components of the mixture increases, the flexibility of the distribution increases too, but, at the same time, there is a loss in terms of computational efficiency because there are more parameters to estimate.

Moreover, the additional information of cluster membership that one may obtain in using this model is redundant as the main purpose is just to operate a dimension reduction from the ordinal space to the continuous space in order to apply the standard classification methods. As we operate in a framework of supervised classification the class membership is known in advance and this information is used only when the classifier is applied. Thus, considering a mixture of distributions may not be the best possible option.

As alternative, we propose to choose a Beta distribution for the latent variable. The Beta distribution appears useful for our task since, as the α and β parameters change, it assumes many different shapes. Using the Beta we

include the case of symmetric latent distribution (for which no significant differences with respect to using a Gaussian latent distribution are expected in terms of classification performance) but we also include possible deviations from symmetry, which would possibly be more appropriate when dealing with data coming from heterogeneous sub-populations. In particular, the Beta distribution covers the case of highly skewed latent distribution, as well as the case of uniform latent distribution (when $\alpha = \beta = 1$).
Therefore, we assume:

$$f(z) = \frac{1}{B(\alpha,\beta)} z^{\alpha-1}(1-z)^{\beta-1}, \qquad z \in [0,1]$$

where $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$ is the Gamma function.
The choice of a latent Beta distribution has been preferred to the Gaussian mixture distribution because, as mentioned, it presents advantages in terms of computational efficiency. This due to the fact that considering a mixture distribution the number of parameters of the model increases. Indeed, for each component of the mixture, mean and variance have to be estimated, together with the corresponding weight, while with the Beta distribution only the α and β parameters have to be estimated.
Table 3.3 shows the computational times (in seconds) needed by the EM algorithm to estimate the model parameters for the two distributional forms. The computational times in the table have been obtained by applying the two methods to a dataset of 5 ordinal variables and 100 observations. The dataset has been obtained by applying the exponential function to observations randomly generated from a standard normal distribution. Ordinal variables have been obtained by subsequently discretizing the observations. In order to obtain comparable results in terms of computational times the parameters of the EM algorithm such as the number of iterations have been held fixed and equal to 5 among the two methods compared. The structure of the EM algorithm will be discussed in the next section.

Table 3.3: *Computational times (in seconds) required for estimating model parameters with 5 iterations of the EM algorithm for different distributions of the latent variable. The reference dataset consists of 5 ordinal variables and 100 observation.*

	Estimation times
Beta distribution	41.03
Gaussian mixture with 2 components	45.72
Gaussian mixture with 3 components	55.57
Gaussian mixture with 4 components	79.4

One may notice that moving from a mixture of multivariate Gaussian distributions to a Beta distribution also the domain changes. The latent variable upon which classification is based does not assumes any more values in $\mathbb{R}$ but in the interval $[0, 1]$. This is not affecting the classification results as the class information is retained using the response function approach because the only information we use is the order among the categories.

In the next two sections methods for parameter estimates in BRFA are introduced together with the Beta factor scores upon which classification is based.

3.4 Model estimation

In this section we present the method used to obtain parameter estimates considering the Beta response function approach.
Denote:

- $\theta_i = (\boldsymbol{\lambda_{i0}}, \lambda_i)$, the vector of thresholds and factor loading for the i^{th} observed variable, where $\boldsymbol{\lambda_{i0}}$ is a $(k_i - 1)$-dimensional vector.
- $\boldsymbol{\delta} = (\alpha, \beta)$, the Beta distribution parameters.
- $\boldsymbol{\tau} = (\boldsymbol{\theta}, \boldsymbol{\delta})$, the model parameters, collectively.

Model parameters can be estimated by the EM algorithm (Dempster *et al.*, 1977), which is an iterative method for calculating the maximum likelihood estimates when we deal with latent variables. Indeed, in the presence of latent variables, the classical maximum likelihood method through the computation of the gradient becomes infeasible. Using the EM algorithm the problem of maximizing the incomplete log-likelihood $\log f(\mathbf{x};\boldsymbol{\tau})$ is reformulated into the problem of maximizing (M-step) the conditional expected value (E-step) of the complete log-likelihood $\log f(\mathbf{x},z;\boldsymbol{\tau})$, which is computationally simpler. Since implementing an EM algorithm for a Beta latent variable presents unique challenges, dedicated algorithms have been developed in the statistical software **R** as no packages are currently available.

The EM algorithm used in this study consists in the following steps:

1. Select initial values for $\tilde{\boldsymbol{\tau}} = (\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\delta}})$.
2. Calculate $\boldsymbol{\tau}$ which maximize $E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log \sum_{h=1}^{n} f(\boldsymbol{x}_h, z_h; \boldsymbol{\tau})\right]$.
   Set $\tilde{\boldsymbol{\tau}} = \boldsymbol{\tau}$.
3. If the change in the observed data log-likelihood is greater than a fixed ϵ, i.e. if convergence is not achieved, return to step 2 an iterate until convergence.

In the first step of the algorithm the initial values of the model parameters have been chosen randomly sampling from a uniform distribution. In particular, the intercept values and the factor loading for each observed variable have been randomly sampled from $Unif(-2,2)$ and $Unif(0,3)$, respectively. The intercept values have been subsequently sorted in order to reflect categories order. The Beta distribution parameters α and β have been sampled at random from $Unif(0,10)$.

The complete log-likelihood can be expressed as:

$$f(\mathbf{x},z) = f(\mathbf{x}|z)f(z)$$

thus, the second step of the algorithm can be rephrased as:

- Calculate $\boldsymbol{\theta}$ which maximize $E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(\mathbf{x}|z;\boldsymbol{\theta})\right]$;
- Calculate $\boldsymbol{\delta}$ which maximize $E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(z;\boldsymbol{\delta})\right]$.

where:

$$E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(\mathbf{x}|z;\boldsymbol{\theta})\right] = \int_0^1 \log f(\mathbf{x}|z;\boldsymbol{\theta}) f(z|\mathbf{x};\tilde{\boldsymbol{\tau}})\, dz \tag{3.9}$$

and

$$E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(z;\boldsymbol{\delta})\right] = \int_0^1 \log f(z;\boldsymbol{\delta}) f(z|\mathbf{x};\tilde{\boldsymbol{\tau}})\, dz \tag{3.10}$$

In order to compute the conditional expected value of the complete log-likelihood it is then necessary to determine the conditional distribution of the latent variable given the observed variables. From Bayes theorem the distribution is given by:

$$\begin{aligned} f(z|\mathbf{x};\tilde{\boldsymbol{\tau}}) = \frac{f(z,\mathbf{x};\tilde{\boldsymbol{\tau}})}{f(\mathbf{x};\tilde{\boldsymbol{\tau}})} &= \frac{f(z;\tilde{\boldsymbol{\delta}})f(\mathbf{x}|z;\tilde{\boldsymbol{\theta}})}{f(\mathbf{x};\tilde{\boldsymbol{\tau}})} \\ &= \frac{f(z;\tilde{\boldsymbol{\delta}})f(\mathbf{x}|z;\tilde{\boldsymbol{\theta}})}{\int_0^1 f(z;\tilde{\boldsymbol{\delta}})f(\mathbf{x}|z;\tilde{\boldsymbol{\theta}})\partial z} \end{aligned}$$

where:

- $f(z;\tilde{\boldsymbol{\delta}})$ is the Beta distribution with known parameters;
- $f(\mathbf{x}|z;\tilde{\boldsymbol{\theta}}) = \prod_{i=1}^{p} f_i(x_i|z,\tilde{\theta}_i)$ is the conditional distribution of the manifest variables with known parameters.

Considering that we express the conditional cumulative probabilities in equation (3.5) as a linear combination of latent scores through a logit link function, the integral at the denominator has the following form:

$$f(\mathbf{x};\tilde{\boldsymbol{\tau}}) = \int_0^1 \frac{1}{B(\alpha,\beta)} z^{\alpha-1}(1-z)^{\beta-1} \prod_{i=1}^{p}\prod_{l=1}^{k_i}\left[\frac{e^{\lambda_{i0(l)}-\lambda_i z}}{1+e^{\lambda_{i0(l)}-\lambda_i z}} - \frac{e^{\lambda_{i0(l-1)}-\lambda_i z}}{1+e^{\lambda_{i0(l-1)}-\lambda_i z}}\right] dz \tag{3.11}$$

Since equation (3.11) cannot be analytically solved it is approximated. Among the several possible approximation methods, Gauss-Legendre quadrature points have been used.

Gauss-Legendre quadrature points approximation

The general formula for approximating an integral with Gaussian quadrature is the following:

$$\int_a^b \omega(x) f(x)\, dx = \sum_{i=1}^{n} w_i f(x_i) + R_n,$$

where $\omega(x)$ is the weight function, x_i are the n nodes (i.e. points on the abscissa axis where the function is evaluated) and w_i are the weights. R_n is the approximation error term.

The Gauss-Legendre integration formula is one of the most used as it has the highest possible precision degree and it is analytically exact for polynomials of degree at most $2n-1$ if nodes correspond to the roots of the orthogonal polynomial for the same $[a, b]$ interval and weighting function (Babolian *et al.*, 2005). The nodes in Gaussian quadrature are always interior points of the reference interval so the Gaussian formulae provides advantages when the function assumes an infinite value at one end of the interval. Several Gaussian quadrature methods exist, which differentiate for the integration interval, weight functions and related orthogonal polynomials. Some of these methods are reported in Table 3.4. Gauss-Legendre quadrature method provides for the case where the integral interval is $[-1, 1]$, with $\omega(x) = 1$. The quadrature points are the zero points of the Legendre polynomials of the first kind $P_n(x)$. Legendre polynomials are sometimes expressed through what is know as Rodrigues' formula (Askey, 2005), introduced independently by Rodrigues (1815), Ivory (1824) and Jacobi (1827):

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^x} \left[(x^2-1)^n\right]$$

Table 3.4: *Some characteristics of Gaussian quadrature methods.*

Interval	$\boldsymbol{\omega}$**(x)**	**Related orthogonal polynomials**
$[-1, 1]$	1	Legendre polynomials
$[-1, 1]$	$1/\sqrt{1-x^2}$	Chebyshev polynomials
$[0, \infty]$	e^{-x}	Laguerre polynomials
$[-\infty, \infty]$	e^{-x^2}	Hermite polynomials

An integral on the interval a and b of a generic continuous function $f(x)$ is approximated through the Gauss-Legendre quadrature method as follows:

$$\int_a^b f(x)\partial x = \int_{-1}^1 f\left(\frac{b-a}{2}x' + \frac{b+a}{2}\right)\frac{b-a}{2}dx' \tag{3.12}$$

$$= \frac{b-a}{2}\sum_{i=1}^n \omega_i f\left(\frac{b-a}{2}x'_i + \frac{b+a}{2}\right) + R_n, \tag{3.13}$$

where $-1 < x' < 1$ and

$$x' = \frac{x - \frac{b-a}{2}}{\frac{b-a}{2}}, \quad x = \frac{b-a}{2}x' + \frac{b+a}{2}, \quad w_i = \frac{2}{(1-x_i'^2)[P_n(x_k')]^2}.$$

In equation (3.3) we have that $a = 0$ and $b = 1$, thus:

$$\begin{aligned} f(\mathbf{x};\tilde{\boldsymbol{\tau}}) &= \int_0^1 f(z;\tilde{\boldsymbol{\tau}})f(\mathbf{x}|z;\tilde{\boldsymbol{\tau}})\,dz \\ &= \frac{1}{2}\int_{-1}^1 f(\frac{z'}{2}+\frac{1}{2};\tilde{\boldsymbol{\tau}})f(\mathbf{x}|\frac{z'}{2}+\frac{1}{2};\tilde{\boldsymbol{\tau}})\,dz' \\ &= \frac{1}{2}\sum_{i=1}^n w_i f\left(\frac{z_i'}{2}+\frac{1}{2}\right) f(\mathbf{x}|\frac{z_i'}{2}+\frac{1}{2};\tilde{\boldsymbol{\tau}}) + R_n. \end{aligned}$$

Therefore, the joint density function of the observed variables has been approximated as the weighted sum of the integrand in equation (3.11), evaluated over a linear transformation of n quadrature points z_i for $i = 1, \ldots, n$. In the simulations we have chosen to use $n = 10$ quadrature points.

Once the approximation of $f(\mathbf{x};\tilde{\boldsymbol{\tau}})$ is obtained, the M-step of the algorithm consists in maximizing (3.9) and (3.10) respectively in $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$.
As an analytical estimator for these parameters cannot be derived, it is necessary to introduce in the M-step of the algorithm a numerical optimization algorithm. This leads to a generalized version of the EM algorithm (GEM, McLachlan & Krishnan, 2007).
Estimates for intercepts and factor loading for each observed variable have been obtained thought the Nelder-Mead algorithm or Simplex method (Nelder & Mead, 1965), which is a popular direct search method due to its robustness and its low overhead in storage and computation (Dennis & Woods, 1987 and Lagarias *et al.*, 1998). Generally, the Nelder-Mead algorithm is designed for minimization of an objective function of several variables. In our case it has been applied on the logarithm in (3.9) with inverted sign. For the i^{th} ordinal variable with k_i categories the domain of the function has k_i dimensions ($k_i - 1$ for the intercepts plus one factor loading).
The method requires only function evaluation and not derivatives of the function. As the name suggests the method works with geometrical objects called simplex. In a general h-dimensional space the simplex consists in $h+1$ points (or vertex) and all their interconnecting line segments, polygonal faces, etc. At each step of the algorithm the worst vertex (i.e. the vertex where the

objective function has the highest values) is rejected and replaced with a new one.

The parameters of the Beta distribution have been estimated using the method introduced by Brent (2013). This method is used to find minima (or maxima) of continuous functions in one variable and it is a combination of the golden ratio search (Kiefer, 1953) and the method introduced by Jarratt (1967), which uses successive parabolic interpolations. The two methods are combined in a way that retains the advantages of both. Indeed, with golden ratio search linear convergence is guaranteed. Processes like successive parabolic interpolation do not always converge but if the objective function has a continuous second derivative which is positive at the minimum then convergence is superlinear, with order at least 1.324.
Starting from the initial values assigned to the Beta parameters on the first step of the EM algorithm $\tilde{\boldsymbol{\delta}} = (\tilde{\alpha}, \tilde{\beta})$, the procedure we adopted consists in two steps:

1. Calculate α which maximize $E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(z; \alpha, \tilde{\beta})\right]$ and set $\tilde{\alpha} = \alpha$;

2. calculate β which maximize $E_{z|\mathbf{x},\tilde{\boldsymbol{\tau}}}\left[\log f(z; \beta, \tilde{\alpha})\right]$ and set $\tilde{\beta} = \beta$.

In order to keep computational times low in the performed simulations, both Nelder-Mead and Brent methods are applied considering a maximum number of iterations equal to 10, which is quite low for these kind of algorithms. Anyway, as shown in the later chapters, this leads to quite satisfying results but in order to increase the computational efficiency alternative optimization methods may be tested. We leave this for future studies.

3.5 Factor scores

As mentioned, the main interest is to locate each unit basing on the response pattern on the latent variable space and then proceed with classification. We move from a p-dimensional space of ordinal variables to a one dimensional space with a single continuous variable where we can apply standard classification methods. This problem has been traditionally treated by determining factor scores.

Once the model parameters have been estimated through the EM algorithm, the aim is to obtain the vector of latent variable scores associated with the observations.

Bartholomew (1984) investigated a method to determine component scores that can be used to scale units on the latent dimension when response variables are binary with logit link function. The posterior density function of $z|\mathbf{x}$ depends on the observed binary variables $\mathbf{x}$ through the q components:

$$z_j = \sum_{i=1}^{p} \lambda_{ij} x_i \qquad j = 1, \dots, q$$

where λ_{ij} are the factor loadings.

Unfortunately, no simple linear function exists to summarize the information contained in the latent variable because $\gamma_{i(l)}$ (for $i = 1, \dots, p$ and $l = 1, \dots, k_i$) is not a linear function of z. As suggested by Bartholomew (1981), if all variables in the model are random, the mean of the posterior distribution of z, given the response patter $\mathbf{x}_h$, can be used to score the response pattern:

$$E(z|\mathbf{x}_h) = \int_0^1 z f(z|\mathbf{x}_h)\, dz \tag{3.14}$$

As equation (3.14) cannot be analytically solved either, Gauss-Legendre quadrature points have been used to approximate the integral with a procedure similar to the one shown above.

In Table 3.5 we present an example of how the response patterns are scored. We just show a sample of 10 response patterns simulated from 5 ordinal variables in the same way of data shown in Table 3.3. Each row in Table 3.5 shows the categories in the response pattern for each of the five ordinal variables. Model parameter estimates associated with these data are presented in Table 3.6.

Table 3.5: *Simulated response patterns with associated scores obtained with response function approach with latent Beta distribution.*

	V1	**V1**	**V3**	**V4**	**V5**	$E(z\|\mathbf{x})$
$\mathbf{x}_1$	4	3	3	4	3	0.4293465
$\mathbf{x}_2$	1	1	1	3	3	0.3711781
$\mathbf{x}_3$	1	1	1	1	1	0.3404442
$\mathbf{x}_4$	4	1	1	1	3	0.3758608
$\mathbf{x}_5$	3	3	3	5	2	0.4293887
$\mathbf{x}_6$	3	2	4	3	3	0.4071265
$\mathbf{x}_7$	4	2	4	2	2	0.4008928
$\mathbf{x}_8$	2	2	1	1	1	0.3518677
$\mathbf{x}_9$	4	1	2	2	1	0.3739162
$\mathbf{x}_{10}$	5	2	3	3	2	0.4164381

Table 3.6: *Estimates of model parameters.*

	V1	**V2**	**V3**	**V4**	**V5**
$\lambda_{0(1)}$	-0.732	-0.225	0.170	-1.084	-0.281
$\lambda_{0(2)}$	-0.443	1.373	0.880	-0.383	0.480
$\lambda_{0(3)}$	0.219	2.898	1.905	0.353	1.347
$\lambda_{0(4)}$	1.460	2.282	2.937	1.742	1.876
$\lambda_{0(5)}$	-	-	-	-	-
λ_1	0.948	0.002	0.884	1.210	1.265
	$\alpha = 4.5$		$\beta = 7.2$		

In the simulations performed in the later chapters the reference scores are the ones obtained with the Beta response function approach.

3.6 Advantages of Beta response function approach

We have seen that response function approach, with respect to the URVA, presents some advantages.

First of all, the analysis in the response function approach is focused on the entire response pattern of the individuals so it can be considered as a full information method. Full information methods have a theoretical advantage over limited information methods, in that they produce more efficient parameter estimates because they use all the information available in the data, while limited information methods only use the low order margins to estimate model parameters.

On the contrary, the URVA is usually a limited information method because estimating the model parameters involves the evaluation of a p-dimensional integral (as in equation (3.2)), which is not computationally feasible. Moreover, the URVA can handle a limited number of items due to the large weight matrix needed for parameter estimation with the generalized least squares method.

Another advantage in using the response function approach is that there is no need to define an underlying distribution for each observed variable as happen in the underlying response variable approach or in the conditional mean scores.

Our approach seems attractive in this particular situation as it allows to overcome the limitations of the scoring methods presented in chapter 2 since the choice of the underlying distribution upon which classification is based is lead-by-data, moreover, as the focus is the entire response pattern, there is no loss of information due to possible correlations among variables.

Furthermore, the innovation we propose in introducing a latent Beta distribution allows to a flexible distributional form that may better approximate the "true" underlying mixture distribution with shorter computational times compared to the ones obtained considering a mixture of multivariate Gaussians due to the fact that only two parameters have to be estimated.

Chapter 4

Classification methods

The purpose of this chapter is to introduce the methods that will be used in the simulations to classify the ordinal data. Each one of the presented methods is designed for supervised classification of variables measured on a continuous scale.
We remark that until now the class membership information has never been used in the presented scoring methods (chapter 2) and in the new proposed method BRFA (chapter 3). We introduce this information when the classifier is applied on the scores.

The supervised classification methods used are: the linear and quadratic discriminant analysis, the naive Bayes classifier, the support vector machine with radial basis kernel and the quantile-based classifier. These methods have been chosen because they are among the most used in the supervised classification context (Bishop, 2006; Hastie *et al.*, 2009).
Dedicated classification algorithms are already implemented in specific packages of the statistical software **R** (R Development Core Team, 2008), which has been chosen to perform the simulations. The linear discriminant analysis and quadratic discriminant analysis are available in the **R** package `MASS` (Venables & Ripley, 2002), support vector machine and naive Bayes are implemented in the **R** package `e1071` (Meyer *et al.*, 2015) and quantile classifier is available in the **R** package `quantileDA` (Hennig & Viroli, 2016b). In the simulations, each scoring method presented in chapter 2 and the response variable approach with Beta latent distribution presented in chapter 3 have been applied to the dataset and then the classifiers have been compared basing on the misclassification rate.

Supervised and unsupervised classification

The supervised classification is used in many areas of science and there is plenty of possible applications:

- Biomedical studies: ex. basing on demographic, diet and clinical measurements for the patients predict whether a patient, hospitalized due to a heart attack, will have a second heart attack;
- Text analysis: ex. on the basis of e-mail records assign messages to directories, finding synonyms, SPAM detection, etc;
- Handwriting recognition;
- Speech recognition.

Usually, in the framework of statistics, pattern recognition, or machine learning we refer to classification as the assignment of objects into one of K known classes or populations $\Pi_1, \ldots, \Pi_K$.
In other words, the main goal in classification is to identify to which of a set of possible classes a new observation belongs. Each new observation $\mathbf{x}_{new} = (x_1, \ldots, x_p)$ is supposed to be a realization of a p-variate random variable, whose probability density corresponds to one of K possible population densities.
A general distinction in classification methods is between supervised and unsupervised classification (usually, we refer to the latter as clustering).
In the supervised classification context the assignment of a class label to a new observation is done on the basis of a training set, which is a set of observations whose class is known in advance. The information in a learning set of labeled observations is used to construct a classifier, i.e. a classification rule, which scope is to discriminate (or separate) the predefined classes as much as possible. Ideally, the sample observations in the training set are representative of the corresponding populations. Basing on the analysis of the training data, one should be able to tune the parameters of an algorithm, which can be used for mapping the new observations.
In clustering no prior information about the classes is given. Broadly speaking, the scope is to select and group homogeneous elements in the data set.
From now on, following the machine learning terminology, we will refer to the observations as instances and to the explanatory variables as features (grouped into a feature vectors). We denote as $\mathbf{X}$ the $N \times p$ data matrix.

The i^{th} column of the data matrix $\mathbf{x}_{0i}$ is a $N \times 1$ feature vector (where N is the total number of instances) while the j^{th} row $\mathbf{x}_{j0}$ corresponds to a $p \times 1$ instance vector.

Parametric and nonparametric classification methods

On the basis of the assumptions we made about the data it is possible to further divide the classification methods into parametric methods and nonparametric methods (Hand, 1997).
In parametric methods we make assumptions on the probability distribution of the data conditionally to each class. In nonparametric methods no prior information about the data is given. In these cases classification is based on the local vicinity of the instance to the class.
Nonparametric methods are usually more flexible than parametric ones as they require fewer assumptions about the underlying population from which data are drawn (Hollander *et al.*, 2013) but they are usually slower than the parametric counterpart. Moreover, if the assumptions for parametric methods are fulfilled, they generally result in better performances with respect to nonparametric methods.
Linear and quadratic discriminant analysis and naive Bayes classifier are examples of parametric methods while support vector machine with radial basis kernel and quantile classifier fall within the definition of nonparametric methods.

Since these classification methods present different properties, it seems appropriate to introduce them briefly before proceeding further with the simulations.

4.1 Linear discriminant analysis (LDA)

Linear discriminant analysis comes as a generalization of Fisher's linear discriminant (Fisher, 1936). It is a linear method of classification as the goal is to subdivide the input space into regions corresponding to the classes by linear decision boundaries. For a two-class problem a linear decision boundary corresponds to a hyperplane that partitions the input space into two sets, one for each class.
Linear decision boundaries arises when we fit linear regression models to

the class indicator variable. In the case of K classes the linear regression model for the k^{th} class variable is $\hat{f}_k(\mathbf{x}) = \hat{\beta}_{k0} + \hat{\boldsymbol{\beta}}_k^T\mathbf{x}$ for $k = 1, \ldots, K$ and $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{pk})$. The decision boundary between classes k and k' is the set of points where $\hat{f}_k(\mathbf{x}) = \hat{f}_{k'}(\mathbf{x})$. This set of points is an hyperplane: $(\hat{\beta}_{k0} - \hat{\beta}_{k'0}) + (\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_{k'})^T\mathbf{x} = 0$.
In linear discriminant analysis each instance is assigned to the class which has the highest posterior probability. If we denote $\Pi = (1, \ldots, K)$ the nominal random variable of possible class labels then the posterior probability for the j^{th} instance to belong to class k is $P(\Pi = k|\mathbf{x}_{j0})$.
Suppose $f_k(\mathbf{x}_{j0})$ is the class-conditional probability density function for the j^{th} instance and π_k the prior probability of class k, with $\sum_{l=1}^K \pi_l = 1$. Using Bayes theorem we can write the posterior probability for class k as:

$$P(\Pi = k|\mathbf{x}_{j0}) = \frac{f_k(\mathbf{x}_{j0})\pi_k}{\sum_{l=1}^K f_l(\mathbf{x}_{j0})\pi_l}. \tag{4.1}$$

In linear discriminant analysis $f_k(\mathbf{x}_{j0})$ is supposed to be the probability density function of a multivariate Gaussian distribution:

$$f_k(\mathbf{x}_{j0}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\dfrac{1}{2}(\mathbf{x}_{j0}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}_{j0}-\mu_k)}, \tag{4.2}$$

moreover, in equation (4.2) it is assumed that the covariance matrices are equal between classes $\Sigma_k = \Sigma$, for $k = 1, \ldots, K$.
Substituting (4.2) into equation (4.1) we have:
(4.3)

$$P(\Pi = k|\mathbf{x}_{j0}) = \frac{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_{j0} - \mu_k)^T\Sigma^{-1}(\mathbf{x}_{j0} - \mu_k)\right)\pi_k}{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \sum_{l=1}^K \exp\left(-\frac{1}{2}(\mathbf{x}_{j0} - \mu_l)^T\Sigma^{-1}(\mathbf{x}_{j0} - \mu_l)\right)\pi_l}.$$

Taking the logarithm of equation (4.3) and keeping parts of the equation which depend on k, this is equivalent to assigning the observation to the class which as the largest value of:

$$\delta_k(\mathbf{x}_{j0}) = (\mathbf{x}_{j0})^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k), \tag{4.4}$$

with $k = 1, \ldots, K$. We often refer to (4.4) as linear discriminant function.
In order to determine the hyperplane separating the classes k and k', we look

at the log ratio of the posterior class probabilities:

(4.5)
$$\log \frac{P(\Pi = k|\mathbf{x}_{j0})}{P(\Pi = k'|\mathbf{x}_{j0})} = \log \frac{f_k(\mathbf{x}_{j0})\pi_k}{f_{k'}(\mathbf{x}_{j0})\pi_{k'}} = \\ = \log \frac{\pi_k}{\pi_{k'}} - \frac{1}{2}(\mu_k - \mu_{k'})^T \Sigma^{-1}(\mu_k - \mu_{k'}) + \mathbf{x}_{j0}^T \Sigma^{-1}(\mu_k - \mu_{k'}).$$

Since the parameters of the Gaussian distributions are unknown they are estimated from the training set. The parameters for the k^{th} class are obtained as:

- $\hat{\mu}_k = \sum_{j \in \Pi_k} \mathbf{x}_{j0}/N_k$, where N_k is the number of instances in the k^{th} class.
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{j \in \Pi_k} (\mathbf{x}_{j0} - \hat{\mu}_k)(\mathbf{x}_{j0} - \hat{\mu}_k)^T/(N-K)$.
- $\hat{\pi}_k = N_k/N$.

These linear decision boundaries are the same for every pair of classes and divide the input space into K regions, labelled according to the classes.
Figure 4.1 shows an example of LDA from Hastie *et al.* (2009), where data come from three classes in $\mathbb{R}^2$. Data are a sample of 30 drawn from three Gaussian distributions with the same covariance matrix and different means. Each data point is labelled according with the true class.
In the way hyperplanes are obtained (see equation (4.5)) the linear discriminant analysis is similar to the logistic regression model. Indeed, in logistic regression we have that:

$$\log \frac{P(\Pi = k|\mathbf{x}_{j0})}{P(\Pi = K|\mathbf{x}_{j0})} = \beta_{k0} + \beta_k^T \mathbf{x}_{j0}, \tag{4.6}$$

while equation (4.5) can also be rephrased as a linear function of the instance as linearity is a consequence of the Gaussian assumption and the common variance assumption:

$$\log \frac{P(\Pi = k|\mathbf{x}_{j0})}{P(\Pi = K|\mathbf{x}_{j0})} = \alpha_{k0} + \alpha_k^T \mathbf{x}_{j0}. \tag{4.7}$$

However, although very similar, coefficients in equation (4.6) and (4.7) are estimated differently as in logistic regression we do not assume any distributional form for the class-conditional densities.

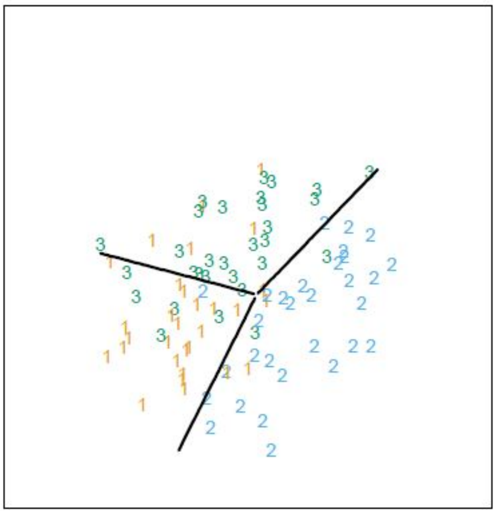

Figure 4.1: *LDA decision boundaries.*

4.2 Quadratic discriminant analysis (QDA)

Quadratic discriminant analysis is closely related to linear discriminant analysis but in this case the class-conditional covariance matrices in equation (4.2) are not assumed to be equal. For this reason, in determining the separating hyperplanes in (4.5), the cancellation of the normalization factor and the quadratic part in the exponents does not occur. This results into a quadratic decision boundaries and a different discrimination function:

$$\delta_k(\mathbf{x}_{j0}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{x}_{j0}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}_{j0}-\mu_k) + \log(\pi_k).$$

As we do not assume equal covariance matrices anymore, they have to be estimated separately for each class, meaning that the number of parameters to estimate is extremely larger than in LDA.

Another way to obtain quadratic decision boundaries is to estimate the parameters as in the LDA case but in the enlarged input space, considering the cross-products and the squares of all the features in the data set. The classification results using this method are usually similar to the ones obtained by simply using the QDA on the original input space.

4.3 Naive Bayes classifier

As in linear and quadratic discriminant analysis, the Bayes classifier assigns each instance to the class having the highest posterior probability:

$$P(\Pi = k|\mathbf{x}_{j0}) = \frac{f_k(\mathbf{x}_{j0})\pi_k}{\sum_{l=1}^{K} f_l(\mathbf{x}_{j0})\pi_l},$$

where $f_k(\mathbf{x}_{j0})$ is the class-conditional probability density function for the j^{th} instance and π_k the prior probability of class k.
For the true but unknown π_k and $f_k(\mathbf{x}_{j0})$, with $k = 1, \dots, K$, the Bayes classifier has the lowest possible error rate. The term naive refers to the fact that this classifier implies two simplifying assumptions on the data (John & Langley, 1995):

1. The features are conditionally independent given the class;

2. No hidden or latent attributes influence the prediction process.

The class-conditional probability density function $f_k(\mathbf{x}_{j0})$ is then given by:

$$f_k(\mathbf{x}_{j0}) = \prod_{i=1}^{p} f_k(x_{ji}).$$

A further assumption that is usually made when dealing with continuous data is that, conditionally to the class, the features are normally distributed. Thus, we have:

$$f_k(\mathbf{x}_{j0}) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(x_{ij}-\mu_k)^2}{2\sigma_k^2}}.$$

As for LDA, parameters estimates are obtained by applying maximum likelihood on the training set. Despite its simplistic form the naive Bayes classifier often provides good results compared with other more complicated classifiers (Langley *et al.*, 1992).

4.4 Support vector machine

The current definition of support vector machine was presented in the work of Cortes & Vapnik (1995).
The hyperplanes described for LDA are found to be optimal (i.e. they create

the biggest margin between the training points for each pair of classes) when the classes are perfectly separated. When the classes overlap and so they may not be separable by a linear boundary, a useful technique, which is a generalization of the linear decision boundaries introduced for LDA is the support vector machine. The support vector machine produces non-linear decision boundaries by constructing linear boundaries in the transformed input space.

Consider a hyperplane separating two classes. In the training set it is assigned value $y_j = 1$ if the j^{th} instance belongs to one class and $y_j = -1$ if the j^{th} instance belongs to the other class.

The p-dimensional linear hyperplane can by described as in the LDA section by:

$$\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp} = 0,$$

if

$$\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp} < 0$$

the j^{th} instance is assigned to class $y_j = -1$ and if

$$\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp} > 0$$

the j^{th} instance is assigned to class $y_j = 1$. Thus, if the classes are separable, we have that:

$$y_j(\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp}) > 0 \quad \forall j = 1, \ldots, p.$$

Hence it is possible to find the estimates of the beta parameters so that the hyperplane creates the biggest margin between the training points for the two classes, thus:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{max} \; M$$

subject to

$$y_j(\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp}) > M \quad \forall j = 1, \ldots, p$$

$$\text{with } \sum_{i=1}^{p} \beta_i^2 = 1.$$

The idea in support vector machines is to create the hyperplanes in a more robust way allowing for instances to be on the wrong side of the boundary. In this way we are able to deal with situations where classes overlap. This is done by considering the slack variables $\xi = (\xi_1, \ldots, \xi_N)$ that represent the proportional amount by which the prediction $\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp}$ is on the wrong side of its margin. Then M is still maximized under the constrictions:

$$y_j(\beta_0 + \beta_1 x_{j1} + \ldots + \beta_p x_{jp}) > M(1 - \xi_j) \quad \forall j = 1, \ldots, p$$
$$\text{with } \sum_{i=1}^{p} \beta_i^2 = 1, \quad \sum_{i=1}^{p} \xi_i = constant.$$

The value of the constant term appear crucial in SVM and has to be tuned on the training set. Indeed, a large value for the constant can cause overfitting while a small value may cause the boundary to be smoother.
Support vector machine described so far is designed for finding linear boundaries for possibly overlapping classes in the input space. An extension of this classifier consists in applying this procedure but in a different input space, obtained considering a transformation of the features through some basis functions $h(\mathbf{x}_{0i}) = (h(x_{1i}), \ldots, h(x_{Ni}))$.
Generally linear boundaries in the enlarged space result into non-linear boundaries in the original space. A very large or infinite dimension of the enlarged space should allow for a better training-class separation but this will possibly result in high computational times. SVM deal with these issues by involving the basis functions in the computations only through the kernel function:

$$K(\mathbf{x}, \mathbf{x}^T) = \langle h(\mathbf{x}), h(\mathbf{x}^T) \rangle.$$

In the simulations we have considered the radial basis kernel:

$$K(\mathbf{x}, \mathbf{x}^T) = exp(-\gamma \|\mathbf{x} - \mathbf{x}^T\|^2).$$

4.5 Quantile classifier

The quantile-based classifier, introduced by Hennig & Viroli (2016a), is a classification method that tries to avoid the problems arising when we deal with potentially high-dimensional data. In these high-dimensional settings often the computations require long time and can be cumbersome. The quantile-based method attempts to overcome these issues by considering a distance-based classifier using only the partial information of the class conditional distributions. Distance-based methods typically consider as class-conditional information the central moments of the distributions (Jörnsten, 2004; Dabney, 2005; Fan & Fan, 2008). In particular, the quantile classifier comes as a generalization of the median-based classifier (Hall *et al.*, 2009), which assigns the instances on the base of the distance from the class-conditional medians.

In the quantile classifier each instance is assigned to a class according to the sum of component-wise distances to the within-class quantiles.

Consider a generic univariate random variable X with cumulative distribution function F_X. Then, the θ^{th} quantile of X, denoted as $q_X(\theta)$, is equal to:

$$q_X(\theta) = F_X^{-1}(\theta) = inf\{x : F_X(x) \geq \theta\}, \quad \theta \in [0,1]$$

The quantile classifier rule allocate a new instance $\mathbf{x}_{j0}$ to the class which gives the lowest quantile distance:

$$\underset{k}{argmin} \sum_{i=1}^{p} \Phi_{ki}(\mathbf{x}_{j0}, \theta) \quad k = 1, \ldots, K$$

where:

- $q_{ki}(\theta)$ is the θ^{th} class-conditional quantile for the j^{th} feature;
- $\Phi_{ki}(\mathbf{x}_{j0}, \theta) = \left(\theta + (1 - 2\theta)\mathbb{1}_{[x_{ji} \leq q(\theta)]}\right) |x_{ji} - q\,(\theta)\,|$ is the quantile distance between x_{ij} and $q_{ki}(\theta)$.

For $\theta = 0.5$ the quantile classifier corresponds to the median classifier.

The optimal value of θ is determined empirically over a grid of values in the interval $[0,1]$ basing on the misclassification rate in the training set and it is unique for all the features. When $p = 1$ and there are only two possible classes, Π_1 and Π_2, the classification rule takes the following simple form:

$$\begin{gathered} if \;\; x_{j0} \leq \tilde{q}(\theta) \;\; then \;\; x_{j0} \in \Pi_1 \\ if \;\; x_{j0} > \tilde{q}(\theta) \;\; then \;\; x_{j0} \in \Pi_2, \end{gathered}$$

where $\tilde{q}(\theta)$ is a cutoff point given by the weighted average of the two class conditional-quantiles. Letting $q^{(1)}(\theta) = min\{q_1(\theta), q_2(\theta)\}$ and $q^{(2)}(\theta) = max\{q_1(\theta), q_2(\theta)\}$ the cutoff is given by:

$$\tilde{q}(\theta) = \theta q^{(1)}(\theta) + (1 - \theta) q^{(2)}(\theta). \tag{4.8}$$

Figure 4.2 shows an example of a two class decision problem in a univariate setting where the populations are distributed according to location shifted χ^2

with 5 degrees of freedom. The first plot shows the case of median classifier (i.e. $\theta = 0.5$) where the cutoff point is the averages between the population medians. The second plot shows the case of optimal θ value for the quantile classifier ($\theta = 0.202$). The error probability region (displayed in gray) is greater for the median classifier respect to the best quantile classifier (for which it reaches the minimum possible value).

Although the quantile classifier is a recently proposed method and, therefore, not among the most used, it has an interesting propriety, which motivates its inclusion among the classification methods considered in this study. For a two class decision problem in the univariate case, considering the optimal theoretical θ value, the cutoff given in equation (4.8) is the optimal decision boundary point that minimizes the overall misclassification probability (we refer the reader to Hennig & Viroli, 2016a for the proof). This property may result useful as we operate classification in the univariate input space obtained through dimension reduction with the response function approach with Beta latent distribution.

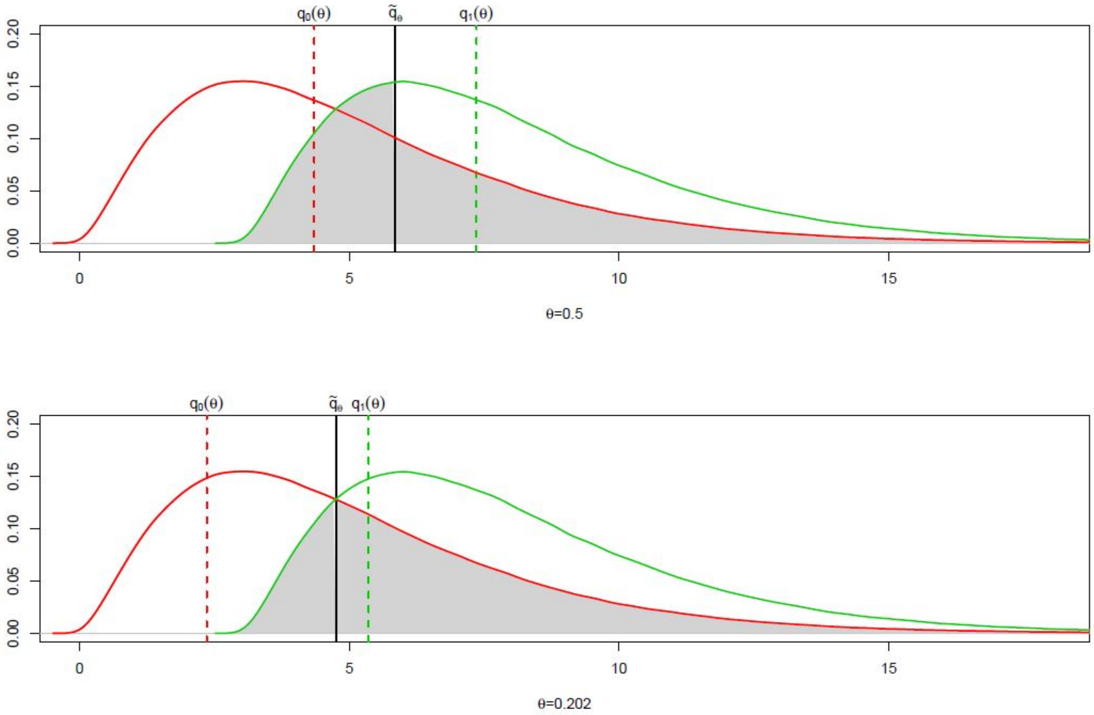

Figure 4.2: *Two location shifted χ^2 distributions and total misclassification probability (in gray) according to: (a) median classifier and (b) the best quantile classifier.*

Chapter 5

Simulation Study

In this chapter the simulation study performed in order to evaluate the performance of the classification methods described in chapter 4 on the scored data is presented. We considered the effect of different factors on the classifiers results, specifically: the number of features (p), the sample size (N) and the number of categories of the ordinal variables (C).
The performance of the classifiers have been evaluated in terms of the mean misclassification rate from a 10-fold cross-validation.
The classification methods used in the simulations are:

- Quantile-based classifier;
- Linear discriminant analysis (LDA);
- Quadratic discriminant analysis (QDA);
- Support vector machine (SVM);
- Naive Bayes classifier (NB).

The scoring methods applied to the ordinal data set are the ones presented in chapter 2, together with the Beta response function approach:

- Raw scores (Raw);
- Ridit scores (Ridit);
- Blom scores (Blom);
- Normal median scores (NM);

- Conditional mean scoring functions:
  - Normal mean scores (NMS);
  - Logistic Mean Scores (LMS);
  - Log-Normal Mean Scores (LNMS).
- Beta response function approach (BRFA).

In the BRFA case the classifier is applied on the vector of latent variable scores, i.e. the expected values of the latent variable, given the response patterns.
Dedicated algorithms have been developed for the scoring methods in the statistical software **R** as no packages are currently available.
All the presented simulation results consider a two class decision problem. The data sets have been generated from two continuous populations that will be denoted as Π_1 and Π_2.
We can distinguish between four steps in the procedure adopted for the simulations:

1. Two data sets have been generated randomly sampling two balanced blocks of observations from two location shifted p-variate continuous distributions, corresponding to the classes;

2. The data sets have been merged and subsequently discretized in order to obtain a single ordinal data set;

3. The scoring methods have been applied to the data sets at step 2;

4. Classification has been performed after subdividing the data sets randomly into into training and test sets (10-fold cross-validation).

The aim of these simulations is not to directly compare the classifiers but, instead, to evaluate the effect of the different scoring method upon the classification results.
In the next two sections the scenarios, i.e. the distributional forms, considered for data generation and the procedure adopted for discretizing the resulting continuous features are described in detail.

5.1 Simulation study scenarios

We generated p vectors for the two populations randomly sampling from two location shifted multivariate distributions in four main scenarios. In each scenario the p vectors have been generated so that they are uncorrelated, considering an identity covariance matrix.
It may seems as a limitation for our method to proceed generating independent features as an advantage of the BRFA is to take into account possible correlations between features, which instead is ignored by the other scoring methods. However, we proceed in this way because when we categorize continuous vectors (with the procedure that will be presented in the next section), we introduce a certain degree of correlation between features. This is known as categorization error (Johnson & Creech, 1983) an occurs when several continuous variables are collapsed into ordinal categories. In these case the measurement errors introduced may be correlated.
When the number of simulated features increases for a fixed sample dimension this effect is more evident.

In the first scenario we considered p variables X_i $(i = 1, \ldots, p)$ distributed accordingly to a Student's t distribution with 3 degrees of freedom for Π_1 and p Student's t-distributed variables shifted by 1 for Π_2.
In the second and third scenarios we considered highly skewed distributions. For Π_1 the p vectors have been generated from a multivariate Gaussian and then transformed using the exponential function (second scenario) and the logarithm function of the absolute value (third scenario). Vectors for Π_2 have been generated from the same distributions shifted by 1.5.
In the fourth scenario, which will be denoted as "mixture" scenario, vectors have been randomly sampled from a multivariate Gaussian for Π_1 and subsequently they have been split in five balanced blocks of different transformations. Again, for Π_2 the vectors come from the same distributional form shifted on the right. The transformations considered in the "mixture" scenario are the exponential, the logarithm of the absolute value, the square and the square root. For the fifth block no transformation occurs.
In each simulation setting the same number of instances has been generated from the two populations.
Table 5.1 summarizes the considered scenarios. For simplicity only the distributions in the univariate case are shown. We denote as X the distribution for

the first population and Y the corresponding distribution for the second population. For the “mixture” scenario the subscripts (from 1 to 5) differentiate the blocks.

Table 5.1: *Simulation scenarios.*

Scenario	**Population 1**	**Population 2**
1	$X \sim t_3$	$Y = X + 1$
2	$X = \exp(W)$, with $W \sim N(0,1)$	$Y = X + 1.5$
3	$X = \log\|W\|$, with $W \sim N(0,1)$	$Y = X + 1.5$
Mixture	$X_1 \sim N(0,1)$ $X_2 = \exp(W)$, with $W \sim N(0,1)$ $X_3 = \log\|W\|$, with $W \sim N(0,1)$ $X_4 = W^2$, with $W \sim N(0,1)$ $X_5 = \sqrt{W}$, with $W \sim N(0,1)$	$Y_1 = X_1 + 1$ $Y_2 = X_2 + 1.5$ $Y_3 = X_3 + 1.5$ $Y_4 = X_4 + 1$ $Y_5 = X_5 + 1$

5.2 Ordinal variables generation

In this section we describe the procedure used in order to obtain an ordinal data set after that the continuous vectors have been generated in the scenarios introduced in the previous section.
The procedure is presented considering a single vector $\mathbf{x}$, generated from the mixture of the two populations with equal weights.
The extension to the multivariate case is given by applying the same procedure to all the variables in the data set.
Discretization is done directly on the mixture vector from the two populations Π_1 and Π_2.
We denote as C the number of categories for the ordinal variable we want to obtain and N the vector length (i.e. sample size).
The feature discretization procedure is the following:

1. The $\mathbf{x}$ vector is sorted;
2. The sorted vector is subdivided into C equi-spaced intervals from the 1^{th} to the 99^{th} percentile;
3. Category i is assigned to the observations falling in the i^{th} interval, for $(i = 1, \ldots, C)$;
4. Observations below the 1^{th} percentile are assigned to category 1, similarly, observations above the 99^{th} percentile are assigned to category C.

We considered the 1^{th} and 99^{th} percentile in step (2) instead of minimum and maximum values in order to avoid possible outliers in defining the intervals. Step (4) of the procedure is included in order to keep under control the vector dimension when data are generated. In this way we consider larger intervals for categories 1 and C, thus the proportions of observations falling in these two categories are higher. Anyway, if N is kept relatively small, the number of observation falling out of the range given by the 1^{th} and 99^{th} percentile is small and the differences in the sample proportions are negligible.

In order to be able to apply the scoring methods, it is necessary that each category appears at least once in the data set since all the procedures are based on the sample proportions. To be sure that in step (2) there are not “empty” intervals once the C percentile-based intervals are defined, C observations are removed randomly from the data and replaced with the same number of observations. Each new observation is assigned to a different interval (i.e. category). In this way the sample dimension N is held fixed and, at the same time, the sample proportions do not vary too much as C is small.

5.3 Results of simulation study

We present now the results of the simulation study. The main interest is to evaluate the impact of differently scored data on the performance of the classifiers, which have been compared in terms of the mean misclassification rates obtained from a 10-fold cross-validation. We considered the effect of different factors on the classification results:

- Number of features in the data set, $p = 10, 20, 40, 100$;
- Sample size, $N = 100, 200, 400$;
- Number of categories of the ordinal variables, $C = 3, 5$.
  Each simulated data set is composed by features having all the same number of categories. We considered also features with different number of categories inside the same data set. In this case the data set has been equally subdivided into features with 3,5 and 7 categories (to identify this case we denote it as $C = 357$).

Because the simulations results are many, for the sake of clarity we decide to not show all of them.

This section is articulated as follows: in the first part the results for the quantile-based classifier are shown for all the scenarios, in the second part results for the other classifiers are shown just for scenario 1. All the other results are included in the appendix.

We choose to proceed considering the quantile-based classifier among the classification methods because, as mentioned in the previous chapter, it partially motivates the choice of the Beta response function approach for treating the ordinal data.

Moreover, since the performance of the classifiers proved to not vary too much with different number of categories for the ordinal features we present

only the case $C = 5$.
The plots that will be shown have all the same structure: each block represent a different number of features in the data set. Inside each block the mean misclassification rate in the test sets from the 10-fold cross-validation is reported for different sample sizes ($N = 100, 200, 400$). Black symbols are used for the mean misclassification rates obtained by applying the classifier on the scoring methods from chapter 2 while the red dots are for the BRFA results. We have excluded from the graphs the results for the normal median scores and blom scores as they are generally very similar to the ones of the other scoring methods. These results, together with the standard errors, will be included in subsequent tables.

5.3.1 Quantile-based classifier simulation results

In this section the simulation results obtained by applying the quantile-based classifier on the scored data are presented. Each of the following plots represents a different scenario from which data are simulated.

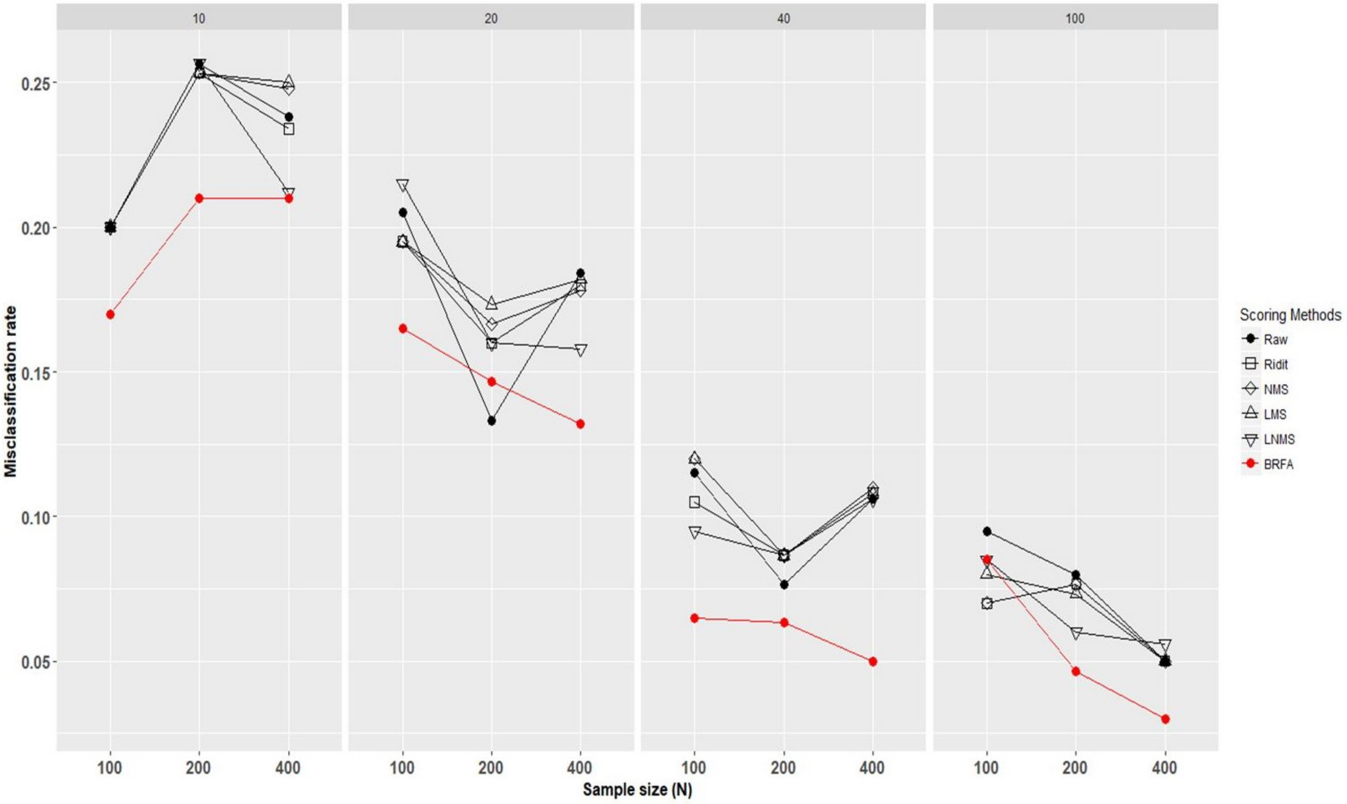

Figure 5.1: *Quantile-based classifier results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

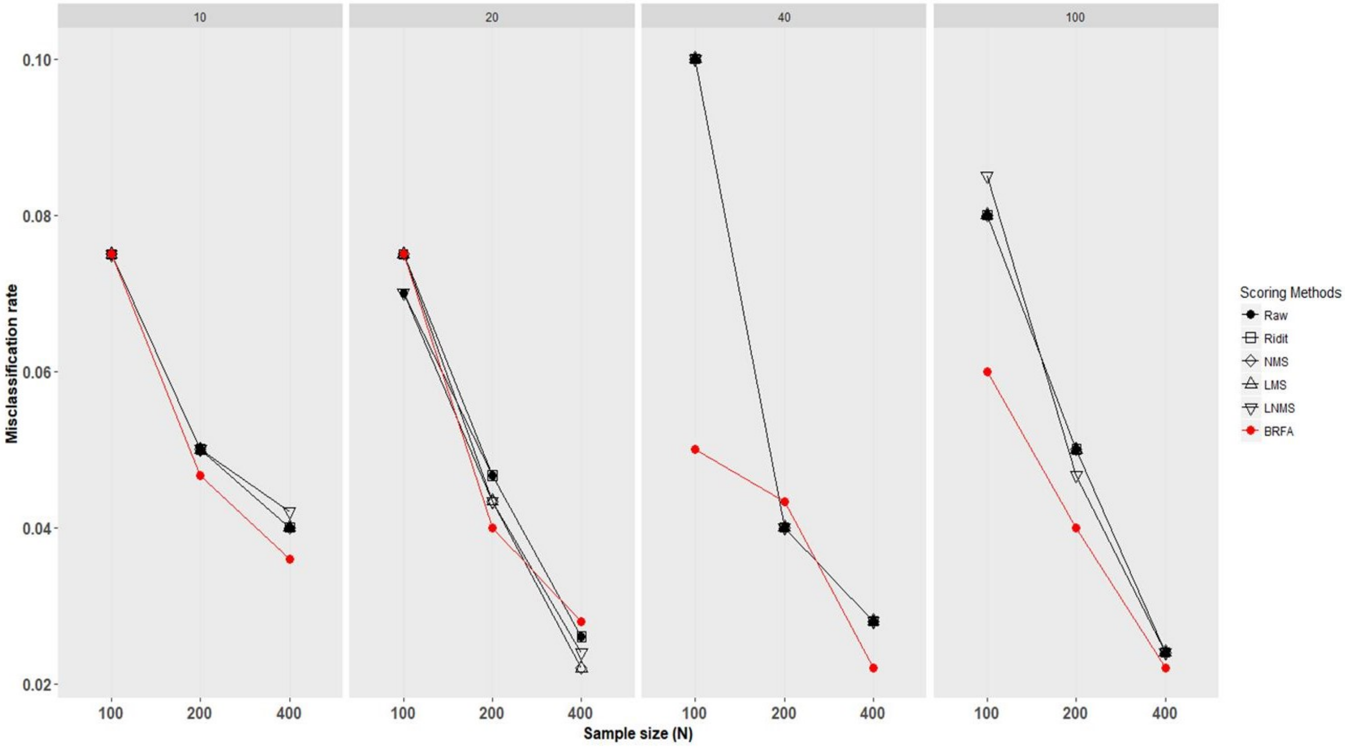

Figure 5.2: *Quantile-based classifier results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

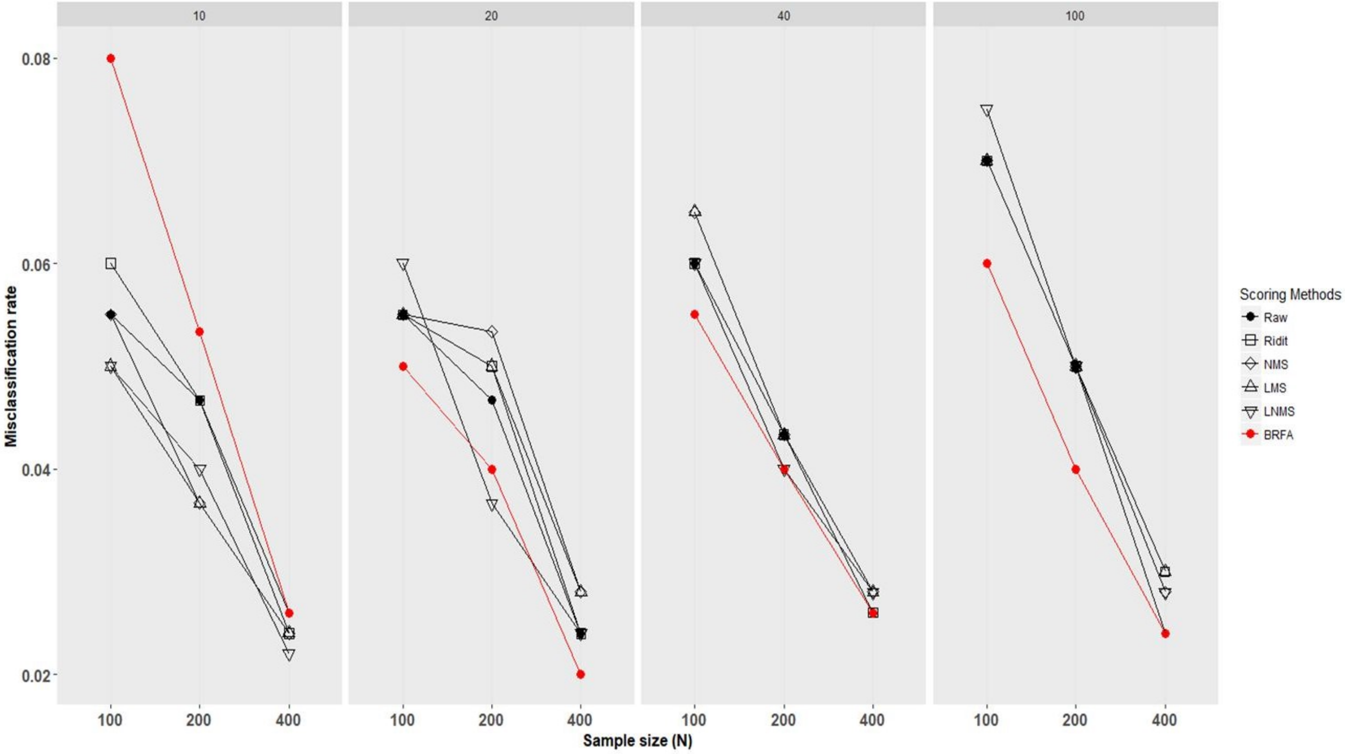

Figure 5.3: *Quantile-based classifier results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

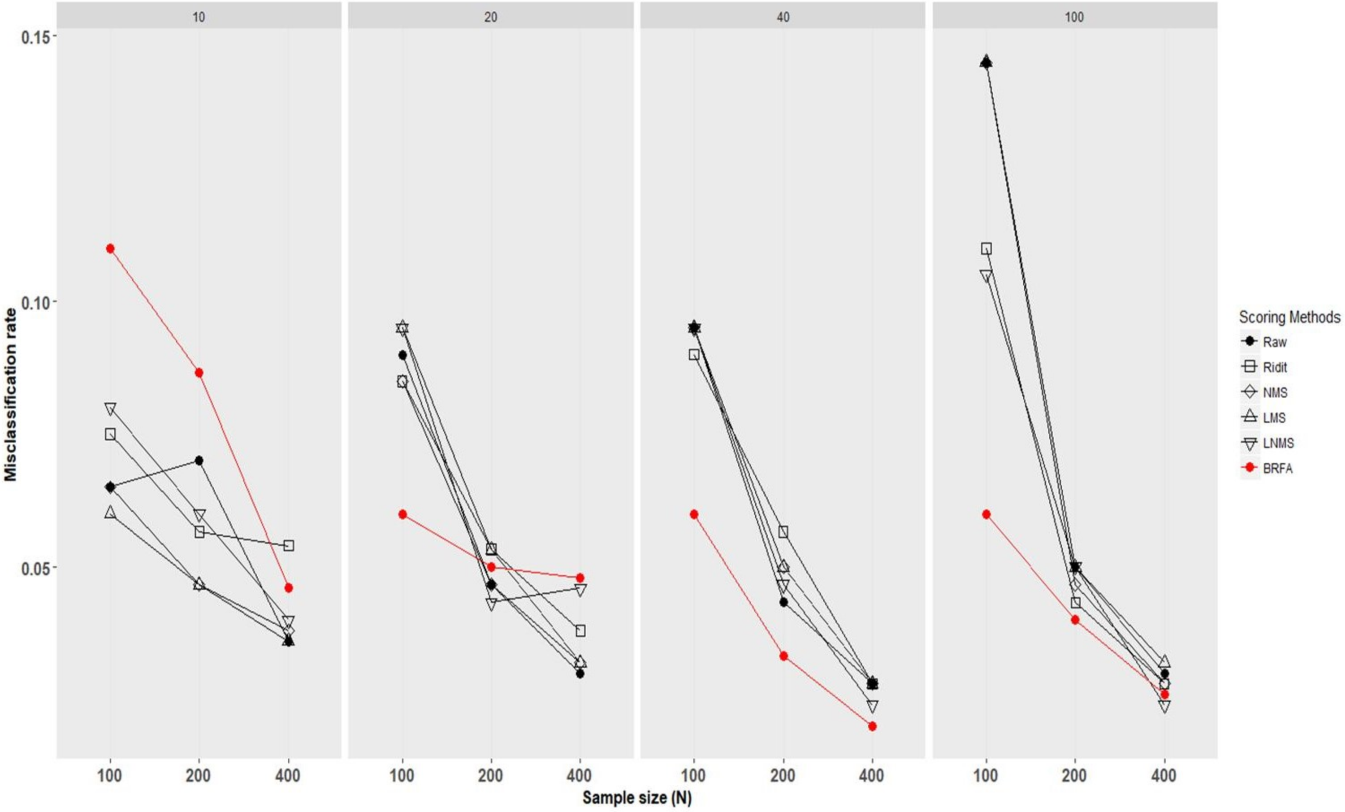

Figure 5.4: *Quantile-based classifier results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

The graphical results shown above are summarized in the following table, together with the standard errors (in brackets). In the table are presented also the results for the normal median scores and blom scores.

Table 5.2: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for quantile-based classifier in the four considered scenarios. Simulation results refer to the case of ordinal features with five categories ($C = 5$).*

Scenario 1												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.20 (0.02)	0.26 (0.03)	0.24 (0.02)	0.20 (0.03)	0.13 (0.01)	0.18 (0.03)	0.12 (0.03)	0.08 (0.01)	0.11 (0.01)	0.1 (0.03)	0.08 (0.01)	0.05 (0.01)
Ridit	0.20 (0.02)	0.25 (0.03)	0.23 (0.02)	0.20 (0.03)	0.16 (0.01)	0.18 (0.03)	0.10 (0.03)	0.09 (0.02)	0.11 (0.01)	0.07 (0.03)	0.08 (0.02)	0.05 (0.01)
NM	0.20 (0.02)	0.25 (0.02)	0.25 (0.02)	0.18 (0.03)	0.16 (0.01)	0.18 (0.03)	0.12 (0.03)	0.09 (0.01)	0.11 (0.01)	0.07 (0.03)	0.07 (0.01)	0.05 (0.01)
Blom	0.20 (0.02)	0.25 (0.02)	0.25 (0.02)	0.18 (0.03)	0.16 (0.01)	0.18 (0.03)	0.12 (0.03)	0.09 (0.01)	0.11 (0.01)	0.07 (0.03)	0.07 (0.01)	0.05 (0.01)
NMS	0.20 (0.02)	0.25 (0.03)	0.25 (0.02)	0.20 (0.03)	0.17 (0.01)	0.18 (0.03)	0.12 (0.03)	0.09 (0.02)	0.11 (0.01)	0.07 (0.03)	0.08 (0.01)	0.05 (0.01)
LMS	0.20 (0.02)	0.25 (0.03)	0.25 (0.02)	0.20 (0.02)	0.17 (0.01)	0.18 (0.03)	0.12 (0.03)	0.09 (0.02)	0.11 (0.01)	0.08 (0.03)	0.07 (0.01)	0.05 (0.01)
LNMS	0.20 (0.02)	0.26 (0.03)	0.21 (0.02)	0.22 (0.03)	0.16 (0.01)	0.16 (0.02)	0.10 (0.03)	0.09 (0.01)	0.11 (0.01)	0.08 (0.03)	0.06 (0.02)	0.06 (0.01)
BRFA	0.17 (0.03)	0.21 (0.03)	0.21 (0.02)	0.16 (0.02)	0.15 (0.02)	0.13 (0.01)	0.06 (0.02)	0.06 (0.02)	0.05 (0.01)	0.08 (0.03)	0.05 (0.01)	0.03 (0.01)
Scenario 2												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.07 (0.01)	0.05 (0.01)	0.03 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
Ridit	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.01)	0.05 (0.01)	0.03 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.01)	0.05 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
Blom	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.01)	0.05 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
NMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.01)	0.04 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
LMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.01)	0.04 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
LNMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.07 (0.01)	0.04 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
BRFA	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
Scenario 3												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
NM	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
Blom	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
NMS	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
LMS	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
LNMS	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.03)	0.05 (0.01)	0.03 (0.01)
BRFA	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
Mixture scenario												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.07 (0.01)	0.04 (0.01)	0.09 (0.02)	0.05 (0.02)	0.03 (0.01)	0.10 (0.02)	0.04 (0.01)	0.03 (0.01)	0.14 (0.03)	0.05 (0.01)	0.03 (0.01)
Ridit	0.08 (0.02)	0.06 (0.01)	0.05 (0.01)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.09 (0.02)	0.06 (0.02)	0.03 (0.01)	0.11 (0.03)	0.04 (0.01)	0.03 (0.01)
NM	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.02)	0.04 (0.01)	0.10 (0.02)	0.05 (0.02)	0.03 (0.01)	0.14 (0.03)	0.05 (0.02)	0.03 (0.01)
Blom	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.02)	0.04 (0.01)	0.10 (0.02)	0.05 (0.02)	0.03 (0.01)	0.14 (0.03)	0.05 (0.02)	0.03 (0.01)
NMS	0.06 (0.03)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.05 (0.02)	0.03 (0.01)	0.10 (0.02)	0.05 (0.02)	0.03 (0.01)	0.14 (0.03)	0.05 (0.02)	0.03 (0.01)
LMS	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.10 (0.03)	0.05 (0.01)	0.03 (0.01)	0.10 (0.02)	0.05 (0.02)	0.03 (0.01)	0.14 (0.03)	0.05 (0.02)	0.03 (0.01)
LNMS	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.10 (0.02)	0.04 (0.01)	0.05 (0.01)	0.10 (0.02)	0.05 (0.01)	0.02 (0.01)	0.10 (0.03)	0.05 (0.01)	0.02 (0.01)
BRFA	0.11 (0.02)	0.09 (0.01)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)

The first thing that can be noticed from the graphs is that, in the majority of cases, the misclassification rates decrease as the sample size increases. This is generally true for every considered classifier. Hennig & Viroli (2016a) have shown that a larger sample size leads to a more consistent choice of the empirical optimal θ and, consequently, an improvement in the performance of the quantile classifier.

Since all the features are informative for the classification task (see the Simulation study scenarios section for this) it is not surprising that for scenario 1 the classifier seems to perform better as p increase for all the scoring systems. This is not true when data come from highly skewed distributions (as for scenarios 2 and 3) or when they come from the mixture scenario. In these cases it seems that there is no gain in adding more features as the discriminant power is already high with $p = 10$. Indeed, the mean misclassification value over all the scoring method, sample sizes and number of feature is 0.051 for scenario 2, 0.044 for scenario 3 and 0.059 for the “mixture” scenario.

From the simulation results it emerges that the performance of the quantile classifier on the Beta factor scores are generally competitive compared to all the other scoring methods.

For scenarios 1 and 2 the quantile applied to the BRFA data provides better or very similar results respect to all others scores for almost all the combinations of N and p.

When the mean misclassification rate obtained with the BRFA is not the smallest one we have that, as can be seen from Table 5.2, the standard errors intervals of the BRFA overlap with those obtained from scoring methods with lower misclassification rates.

Classifying the observations on the base of BRFA appears to be a preferable solution with respect to the approach that in chapter 1 has been called “parametric”, i.e. applying the classifier to the raw scores. On the contrary, it does not seems that there are relevant differences between the error rates associated with raw scores and the error rates obtained through ridits and conditional mean scores. In comparing these approaches no clear patterns emerge and none of them appear to be uniformly better than the others.

For the BRFA method there is almost always an increment in terms of percentage of correctly classified instances, particularly in the first scenario, where data come as a discretization of roughly normal underlying distributions. Unfortunately, this is not always true: in scenario 3 and ‘mixture” scenario we observe situations, when $p = 10$ or $p = 20$, where it is preferable to apply the classifier to datasets obtained by applying alternative scoring methods.

However, it is important to notice that in high dimensional cases ($p = 40$ or $p = 100$) throughout the scenarios there is a gain in applying the quantile on the BRFA, especially when the sample size is $N = 100$ or $N = 200$. As it is possible to see from the graphs in the appendix this often happens also considering the other classification methods.

A possible explanation is that, as mentioned in the first section (Simulation study scenarios), when continuous vectors are discretized, even if the observations are sampled from uncorrelated variables, we introduce a certain degree of correlation between features, which is more evident when the number of features increases for a fixed sample dimension.

When data are scored with ridits or conditional mean scores the correlation between features is not taken into account (as they are applied separately to each feature in the dataset), so there is a loss of information. The situation is different for the BRFA, where the dependence structure of the data is taken into account. Indeed, in the BRFA scores are obtained conditionally to the complete p-dimensional response pattern.

5.3.2 Other classifiers simulation results

This section presents the simulation results obtained by applying the linear and quadratic discriminant analysis, the support vector machine and the naive Bayes classifier on the scored data generated in the first scenario. The following plots show the mean misclassification rate obtained for each combination of p and N. The plots structure is the same presented in the previous section. Because LDA and QDA require the number of predictor variables to be less than the sample size in the case $p = 100$ and $N = 100$ the results for these two classifiers are missing and not displayed in the graphs.

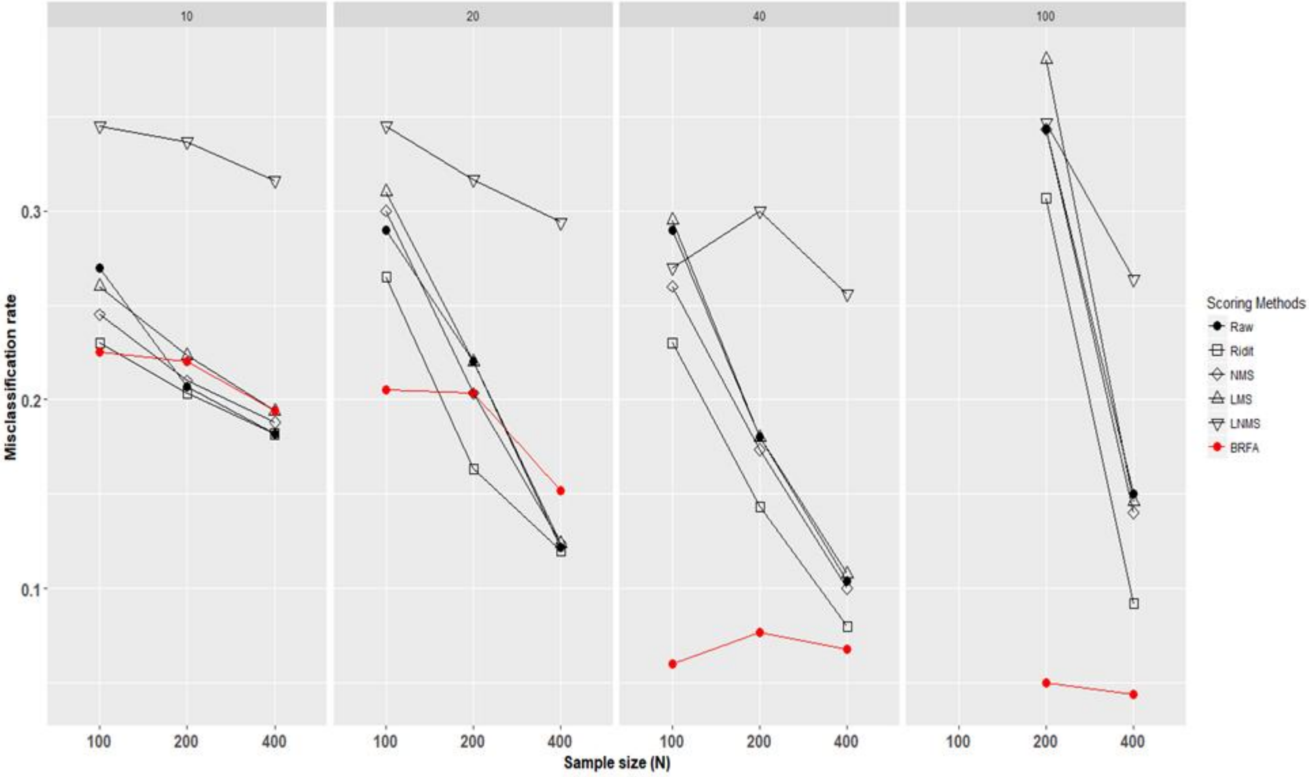

Figure 5.5: *LDA classifier results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

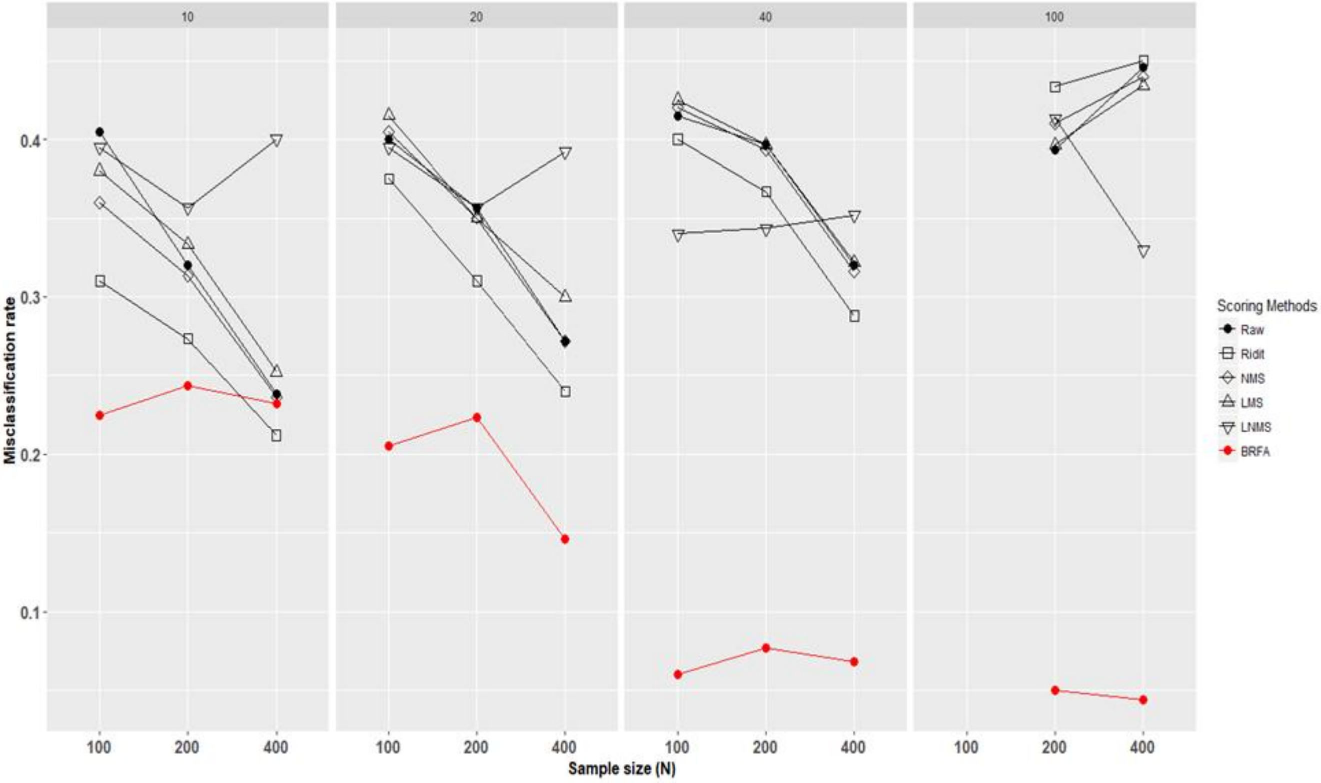

Figure 5.6: *QDA classifier results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

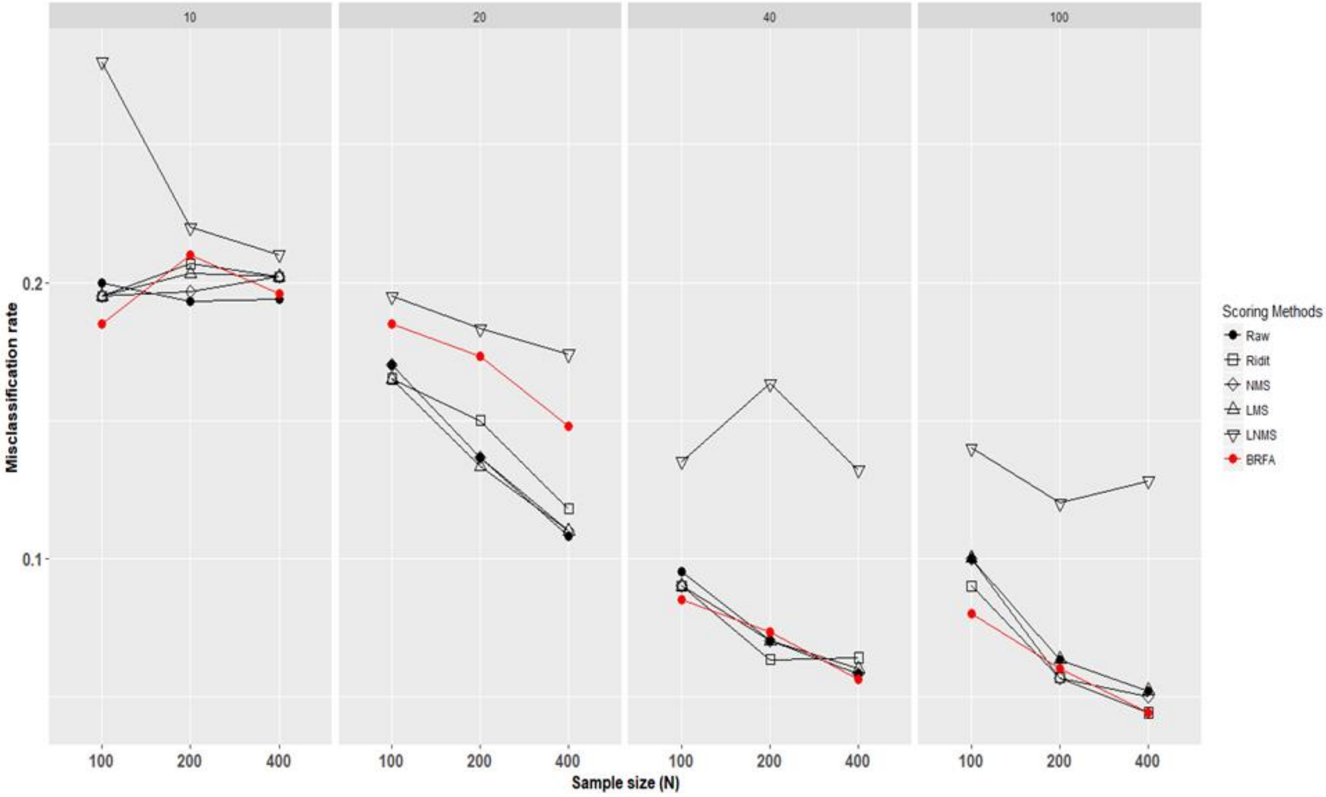

Figure 5.7: *SVM classifier results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

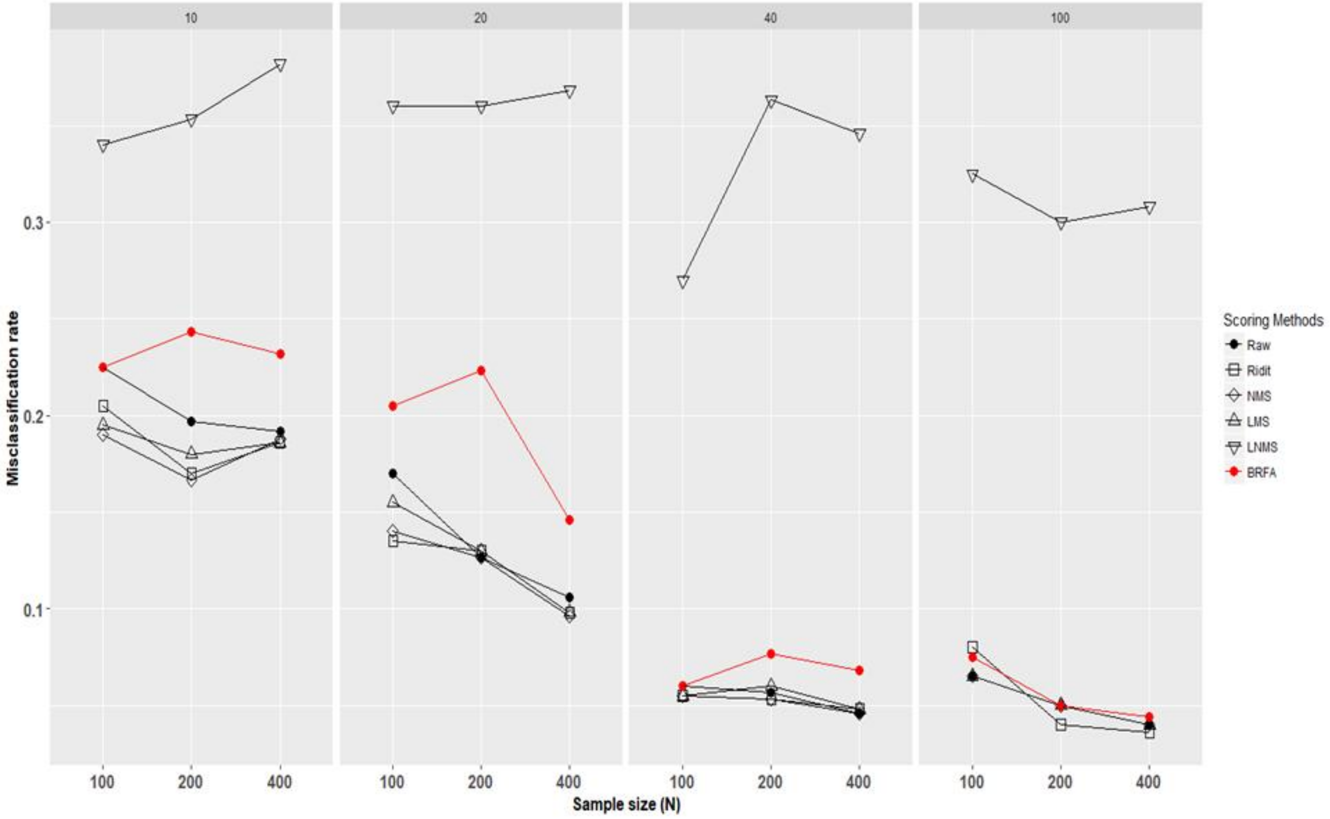

Figure 5.8: *Naive Bayes classifier results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

From Figure 5.5 to Figure 5.8 we notice that, as for the quantile-based classifier results, in the majority of cases the misclassification rates decrease as the sample size and/or the number of features increases in each box of the plots.
A special case is that of log-normal mean scores results associated to each classifier. In the data sets obtained through this scoring system the performance are considerably worse (particularly for the naive Bayes classifier and the support vector machine). Indeed, this type of scoring method is the only one that involve a highly skewed theoretical underlying distribution. Figure

2.3 in chapter 2 shows that log-normal mean scores have smaller distances between lower order categories and bigger distances for higher order categories. In the first scenario, where the data come from an underlying symmetric distribution, this introduces a disturbing element in the allocation method for the classes.

Contrary to what was expected, in scenario 2 (see related graphs in the appendix) the results associated with LNMS are rather poor, in particular when the number of features is low (i.e. $p = 10$) and so the classifier's discriminant power is lower. Conversely, there is a remarkable improvement in the results for LNMS in scenario 3, where the misclassification rate is comparable (and in some cases smaller) to the one obtained with the other scoring methods.

These results are explained by the fact that, broadly speaking, using the conditional mean score methods we are going to modify the distance between categories on the basis of a theoretical latent distribution. Features in scenario 2 are generated from a positively skewed log-normal distribution. Therefore, when we discretize the features we will have a greater number of observations in the lower order categories. As the features are generated from two location shifted distributions with the same functional form, in scenario 2 it is likely that lower order categories bring more information about the classes, viceversa in scenario 3 (as the underlying continuous distributions are negatively skewed) higher order categories are likely to be more discriminative. Considering this, it can be explained why better results are obtained in scenario 3, where LNMS assigns greater distance to higher order categories and, therefore, there is an improvement in the performance of the classification methods.

With the exception of the log-normal mean scores case, it can be noticed that LDA and QDA have worse performance than other classifiers in the first scenario. This is also true for the other scenarios considered (see the relative graphs in the appendix). As mentioned in chapter 4, these classifiers assume linearly separable classes while the data has been simulated so that the classes are overlapped (see simulation study scenarios section).

For what concerns the support vector machines and the naive Bayes classifier results are very good in almost all the considered scenarios and better than those obtained with the linear and quadratic discriminant analysis (particularly for $p = 40$ and $p = 100$). Since the naive Bayes classifier is structurally very similar to the LDA, in that both classifiers assume Gaussian within-class

distributions, such a difference in performance might be surprising.
However, naive Bayes classifier relies on a slightly different distributional model in that it assumes zero off-diagonal covariance (i.e. it assumes that variables within a class are not correlated). Since data in scenarios are generated from independent continuous variables this assumption is totally justified. Thus, for the classification task, ignoring the information given by the empirical correlations between features could improve the results.

Unfortunately, it seems that there is no improvement in terms of misclassification rate in using the SVM or naive Bayes classifier on data scored through the BRFA with respect to all the other scoring methods. For the naive Bayes classifier in $p = 10$ or $p = 20$ cases the results get worse for BRFA. However, as for the quantile-based classifier, in the high dimensional settings there are not significant differences between the scoring methods.

The situation is completely different for linear and quadratic discriminant analysis. Although these classification methods are not particularly suitable for this type of data, the gain in terms of correctly classified cases, if the classifiers are applied on Beta latent variable scores (with particular reference to the cases $p = 40$ and $p = 100$), is really considerable. Also, the classification results for BRFA become comparable to those obtained with naive Bayes classifier, support vector machine and quantile-based classifier.

5.4 Beta response function approach: further advantages

In high dimensional cases all the considered classifiers applied to the observations scored with the BRFA provide better results, or at least comparable, with those obtained for the other scoring methods.
The results obtained with BRFA, although not always the best ones, are surprisingly good considering that the approach is based on a one-dimensional latent variable. Each p-dimensional response pattern is reduced to a single value and this could cause a certain loss of information from the data.
Despite it is precisely the scope of the latent variable model to summarize the information carried from a wider set of observable variables, considering

a single latent factor may be restrictive in some situations. However, from the simulations emerged that, even operating on a one-dimensional space we could have substantial improvements in the classification results. This may also be explained by the fact that with the BRFA we summarise the dependence structure between the ordinal variables. When the features are dependent, the advantage brought by operating with a method that takes into account the dependence structure overcomes the possible loss of information. In the simulated dataset there are two sources of dependence. One source is originated from the class structure, i.e. from how data are simulated we can assume that specific response patterns are likely to be associated with one class and viceversa. The other source of dependence is due to the fact that we collapse the continuous mixture vector from the two populations into ordinal categories. In doing this, correlation among features is introduced.
It will be the subject of future studies to evaluate the effect of more than one latent variable in the BRFA on the classification results.

A further advantage that emerged in the use of BRFA for high dimensional cases is that, unlike the other scoring methods, the misclassification rates obtained with each classifier are not significantly different, regardless to the scenario from which data are generated.
For LDA, QDA and naive Bayes classifier we notice that when they are applied on the single continuous variable obtained through the BRFA the performance are likely to be very similar as they just differ on the assumptions made for the class-conditional variance of the single feature σ_k, for $k = 1, 2$. This appears evident by observing figures from 5.9 to 5.12 reported below. In the plots the mean misclassification rates obtained through the different classifiers for each scoring method (abscissa axis) are shown. The case $p = 100$ and $N = 200$ is considered. Lines are differently dashed and have different colours to distinguish among classification methods. Each plot shows the results for a different scenario.
Differences in misclassification rates are not particularly evident for SVM and naive Bayes classifier, where the results are quite similar between the scoring methods (with the exception of LNMS), while for the quantile-based the performance are better on BRFA.
Instead, considering the linear and quadratic discriminant analysis there is a clear advantage with BRFA. In these cases we get results comparable to those of the other classifiers with an evident gain in terms of misclassification rate,

in particular for the QDA, which is unsuitable for ordinal categorical data (with a percentage of misclassified observations smaller than about 30−40%).

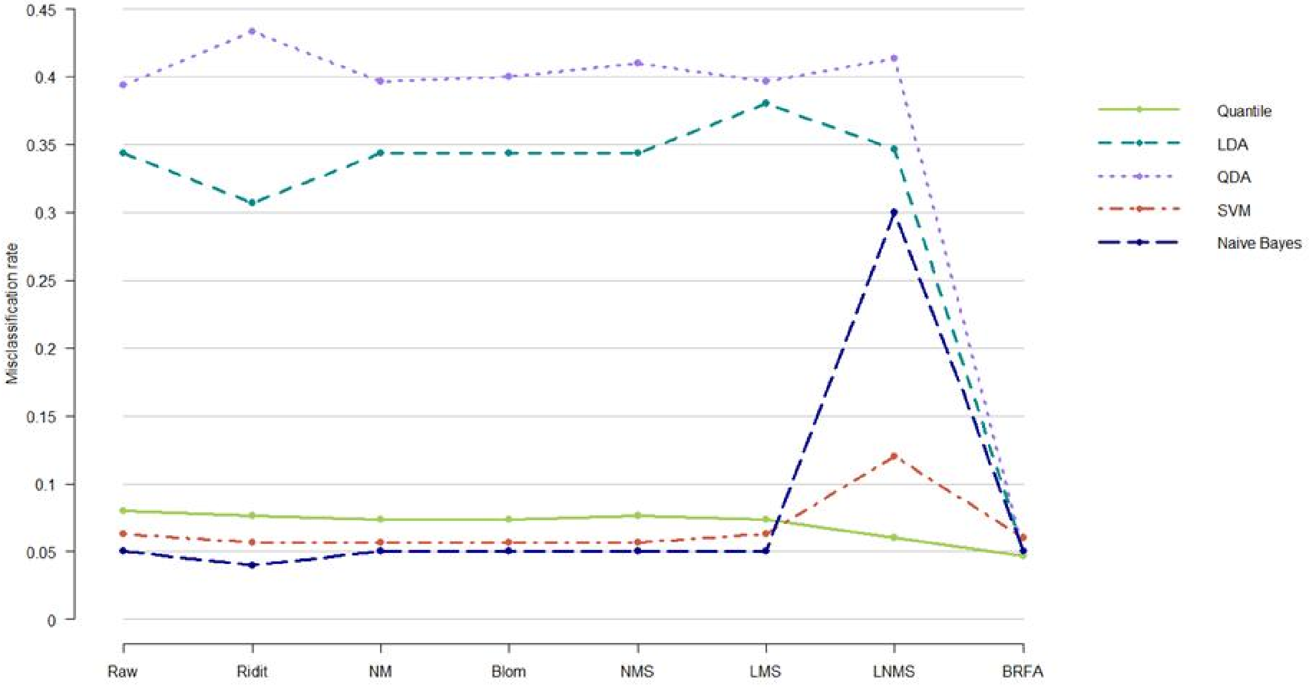

Figure 5.9: *Comparison of classifiers on the scoring methods for* $p = 100$ *and* $N = 200$*. Ordinal variables have been generated from scenario 1 so that they have* $C = 5$ *categories.*

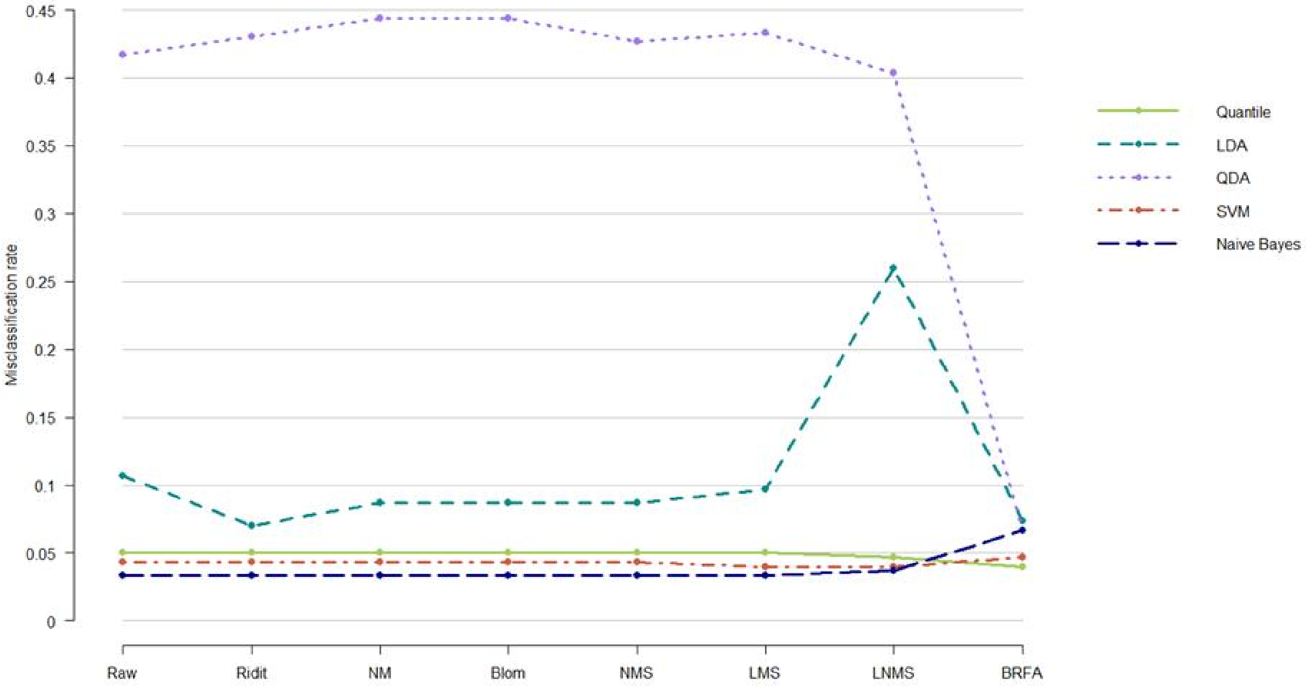

Figure 5.10: *Comparison of classifiers on the scoring methods for $p = 100$ and $N = 200$. Ordinal variables have been generated from scenario 2 so that they have $C = 5$ categories.*

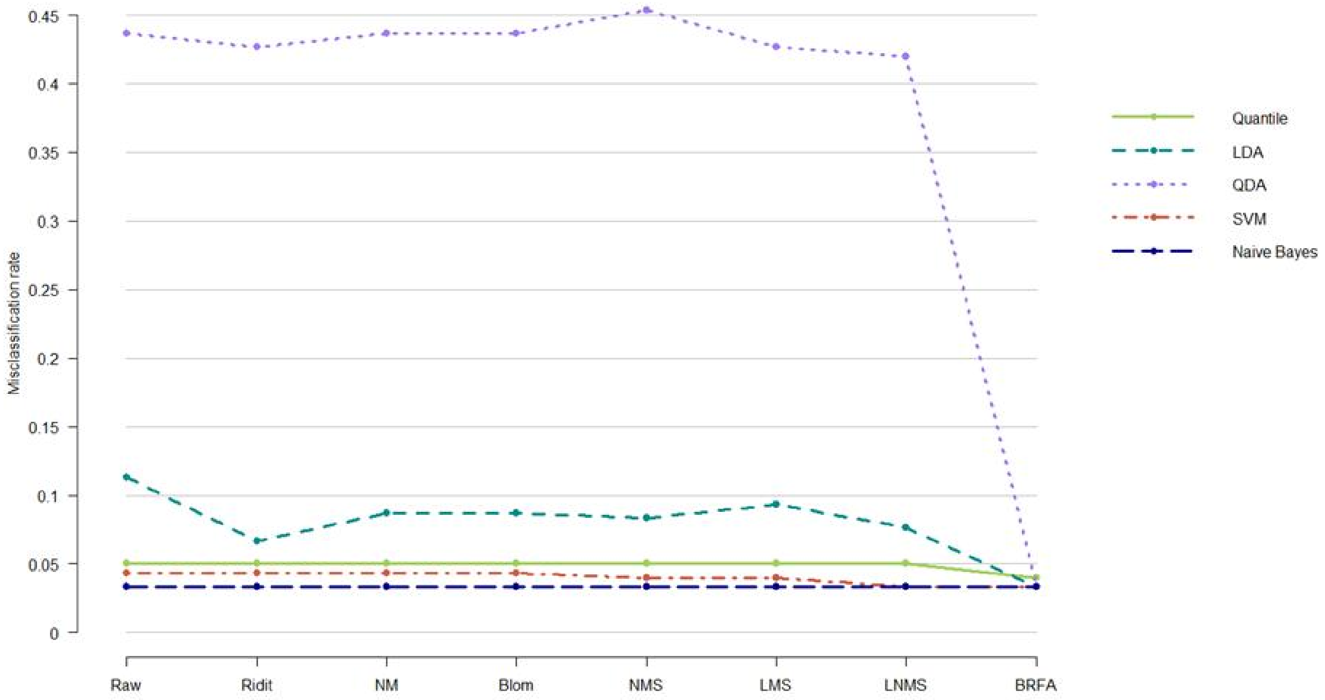

Figure 5.11: *Comparison of classifiers on the scoring methods for* $p = 100$ *and* $N = 200$*. Ordinal variables have been generated from scenario 3 so that they have* $C = 5$ *categories.*

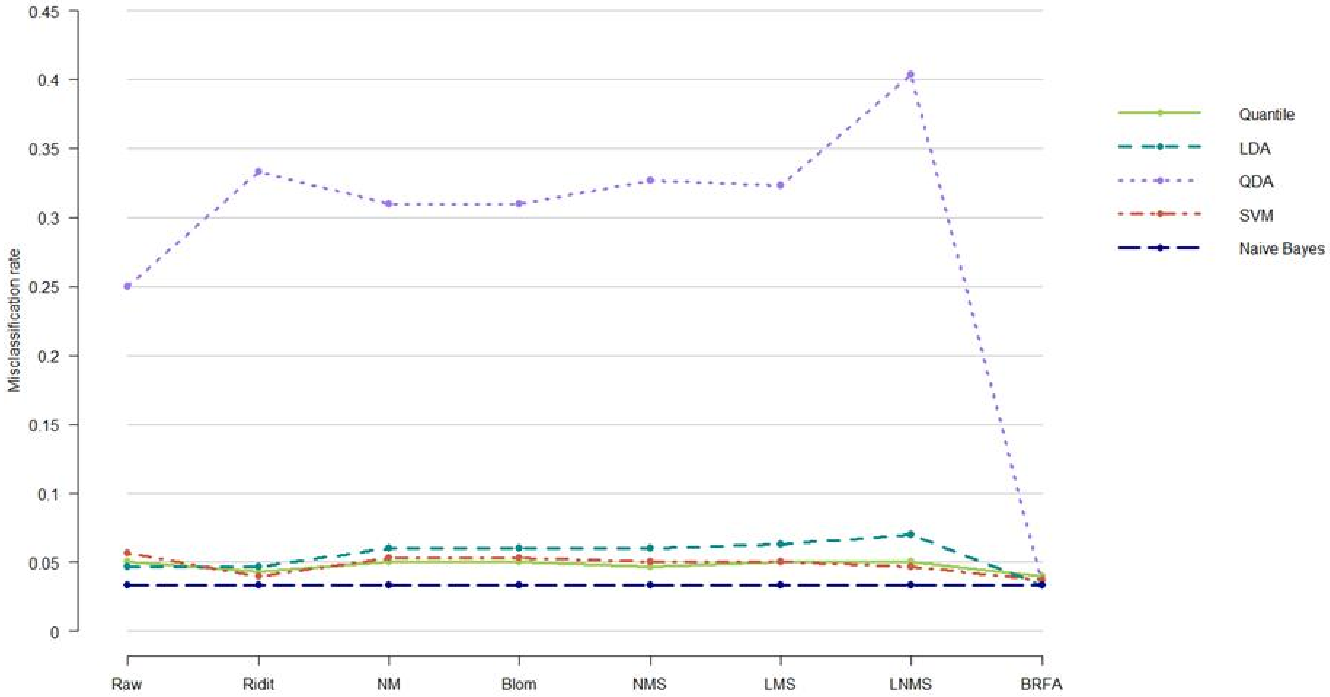

Figure 5.12: *Comparison of classifiers on the scoring methods for* $p = 100$ *and* $N = 200$. *Ordinal variables have been generated from "mixture" scenario so that they have* $C = 5$ *categories.*

Chapter 6

A comparison on a real dataset

In this chapter we compare the classification methods considered in chapter 4 on a real dataset, scored through the scoring methods presented in chapter 2 and the Beta response function approach.
The data have been obtained from the KEEL (Knowledge Extraction based on Evolutionary Learning) data set repository (Alcalá-Fdez *et al.*, 2011). The same dataset is also available from the UCI machine learning repository (Lichman, 2013).
The dataset has been used in the second edition of the Computational intelligence and Learning (CoIL) competition challenge in the year 2000, organized by CoIL cluster. The CoIL cluster is a cooperation between EU funded Networks of excellence with the aim to promote and develop joint cluster internationalisation strategies in different business sectors.
The dataset has been donated by Peter van der Putten of the Dutch data mining company Sentient Machine Research (Van Der Putten & van Someren, 2000) and contains information on customers of an insurance company.
The competition resulted in a wide variety of solutions in terms of approaches and performance and consisted in two main tasks:

1. Predict which customers are potentially interested in a caravan insurance policy (prediction task).

2. Describe the actual or potential customers and possibly explain why these customers buy a caravan policy (description task).

In the application of the classification methods we only focus on the prediction task, but we just report the overall misclassification rate as done in the simulation study.

The dataset proposed in this competition has properties that often appear in real world problems. Features are highly skewed, noisy and correlated. Moreover, there is weak relation between input and target features (Van Der Putten & Van Someren, 2004).

6.1 CoIL dataset description

The dataset consists of 86 features that can be divided in socio-demographic (features 1-43), product ownership (features 44-86) and one binary target feature named “CARAVAN”, which takes values 1 or 0 respectively if the customer is interested or not in the insurance product. Data have been collected from 9822 customers. From the whole set of 9822 customers only 586 resulted to be interested in the insurance product.

In Table 6.1 all the features composing the original data set, along with their description and range of possible values, are reported.

Socio-demographic data are derived from zip area codes so they are linked to the postal code of the customer rather than to the individual customer. All customers living in areas with the same zip code have the same socio-demographic attributes. Because these features are linked to a single hidden variable, i.e. geography, they may be highly correlated.

Since data that refer to socio-demographic have been collected from various and possibly conflicting sources, the measurement noise in the dataset is also high.

For what concerns the product ownership data, the majority of these features are highly skewed with over the 90% of instances falling in the first category. In addition, the average information gain of all features and also of the five most predictive features is very low if compared to some selected datasets from the well know UCI dataset repository (Van Der Putten & Van Someren, 2004).

Table 6.1: *CoIL challenge 2000 data dictionary.*

ID	Name	Description	Domain
1	MOSTYPE	Customer Subtype	[1, 41]
2	MAANTHUI	Number of houses	[1, 10]
3	MGEMOMV	Average size household	[1, 6]
4	MGEMLEEF	Average age	[1, 6]
5	MOSHOOFD	Customer main type	[1, 10]
6	MGODRK	Roman catholic	[0, 9]
7	MGODPR	Protestant	[0, 9]
8	MGODOV	Other religion	[0, 5]
9	MGODGE	No religion	[0, 9]
10	MRELGE	Married	[0, 9]
11	MRELSA	Living together	[0, 7]
12	MRELOV	Other relation	[0, 9]
13	MFALLEEN	Singles	[0, 9]
14	MFGEKIND	Household without children	[0, 9]
15	MFWEKIND	Household with children	[0, 9]
16	MOPLHOOG	High level education	[0, 9]
17	MOPLMIDD	Medium level education	[0, 9]
18	MOPLLAAG	Lower level education	[0, 9]
19	MBERHOOG	High status	[0, 9]
20	MBERZELF	Entrepreneur	[0, 5]
21	MBERBOER	Farmer	[0, 9]
22	MBERMIDD	Middle management	[0, 9]
23	MBERARBG	Skilled labourers	[0, 9]
24	MBERARBO	Unskilled labourers	[0, 9]
25	MSKA	Social class A	[0, 9]
26	MSKB1	Social class B1	[0, 9]
27	MSKB2	Social class B2	[0, 9]
28	MSKC	Social class C	[0, 9]
29	MSKD	Social class D	[0, 9]
30	MHHUUR	Rented house	[0, 9]
31	MHKOOP	Home owners	[0, 9]
32	MAUT1	1 car	[0, 9]
33	MAUT2	2 cars	[0, 9]
34	MAUT0	No car	[0, 9]
35	MZFONDS	National Health Service	[0, 9]
36	MZPART	Private health insurance	[0, 9]
37	MINKM30	Income <30000	[0, 9]
38	MINK3045	Income 30-45.000	[0, 9]
39	MINK4575	Income 45-75.000	[0, 9]
40	MINK7512	Income 75-122.000	[0, 9]
41	MINK123M	Income >123.000	[0, 9]
42	MINKGEM	Average income	[0, 9]
43	MKOOPKLA	Purchasing power class	[1, 8]
44	PWAPART	Contribution private third party insurance	[0, 3]
45	PWABEDR	Contribution third party insurance (firms)	[0, 6]
46	PWALAND	Contribution third party insurane (agriculture)	[0, 4]
47	PPERSAUT	Contribution car policies	[0, 9]
48	PBESAUT	Contribution delivery van policies	[0, 7]
49	PMOTSCO	Contribution motorcycle/scooter policies	[0, 7]
50	PVRAAUT	Contribution lorry policies	[0, 9]
51	PAANHANG	Contribution trailer policies	[0, 5]
52	PTRACTOR	Contribution tractor policies	[0, 7]
53	PWERKT	Contribution agricultural machines policies	[0, 6]
54	PBROM	Contribution moped policies	[0, 6]
55	PLEVEN	Contribution life insurances	[0, 9]
56	PPERSONG	Contribution private accident insurance policies	[0, 6]
57	PGEZONG	Contribution family accidents insurance policies	[0, 3]
58	PWAOREG	Contribution disability insurance policies	[0, 7]
59	PBRAND	Contribution fire policies	[0, 8]
60	PZEILPL	Contribution surfboard policies	[0, 3]
61	PPLEZIER	Contribution boat policies	[0, 6]
62	PFIETS	Contribution bicycle policies	[0, 1]
63	PINBOED	Contribution property insurance policies	[0, 6]
64	PBYSTAND	Contribution social security insurance policies	[0, 5]
65	AWAPART	Number of private third party insurance	[0, 2]
66	AWABEDR	Number of third party insurance (firms)	[0, 5]
67	AWALAND	Number of third party insurane (agriculture)	[0, 1]
68	APERSAUT	Number of car policies	[0, 12]
69	ABESAUT	Number of delivery van policies	[0, 5]
70	AMOTSCO	Number of motorcycle/scooter policies	[0, 8]
71	AVRAAUT	Number of lorry policies	[0, 4]
72	AAANHANG	Number of trailer policies	[0, 3]
73	ATRACTOR	Number of tractor policies	[0, 6]
74	AWERKT	Number of agricultural machines policies	[0, 6]
75	ABROM	Number of moped policies	[0, 3]
76	ALEVEN	Number of life insurances	[0, 8]
77	APERSONG	Number of private accident insurance policies	[0, 1]
78	AGEZONG	Number of family accidents insurance policies	[0, 1]
79	AWAOREG	Number of disability insurance policies	[0, 2]
80	ABRAND	Number of fire policies	[0, 7]
81	AZEILPL	Number of surfboard policies	[0, 1]
82	APLEZIER	Number of boat policies	[0, 2]
83	AFIETS	Number of bicycle policies	[0, 4]
84	AINBOED	Number of property insurance policies	[0, 2]
85	ABYSTAND	Number of social security insurance policies	[0, 2]
86	CARAVAN	Number of mobile home policies	[0, 1]

6.2 CoIL dataset results

Before applying the classification methods we have removed from the dataset all the features that are not measured on an ordinal scale (as the MOSTYPE and MOSHOOFD features) and all the binary features. Classification results refer to the set of the remaining 79 features. As one may notice from Table 6.1, some features are ordinal while other are just integers (as the number of houses, MAANTHUI). The latter have been considered as ordinal features in the computations of scores. In doing this we lose some information as new distances between categories, based on an underlying distribution, are introduced. Indeed, for these features the raw scores are the most appropriate as categories are equi-spaced.

Table 6.2 shows the mean misclassification rates to the whole set of 9822 observations from a 10-fold cross-validation, obtained from each classifier. The columns in the table refer to the classification method considered while the scoring method is reported by row. The standard errors are displayed in brackets.

The performance of linear discriminant analysis and support vector machine appear to be uniformly better with respect to quantile and naive Bayes classifier.

For linear discriminant analysis and support vector machine the classification results are almost identical over the different scoring methods. Instead, we notice that applying the quantile or the naive Bayes on the data scored through the BRFA the misclassification rate is reduced by about the half. In this latter case the classification results are the same for all the classifiers.

Thus, accordingly to what seen in the simulation study, in this example of highly dimensional and correlated data with the Beta response function approach the best classification result is reached independently from the classification method considered.

A particular case is given by the quadratic discriminant analysis. Indeed, applying the QDA on the whole set of features results into a computation error because, as mentioned, these features are highly correlated. Hence, at least one of the covariance matrices can not be inverted and it is not possible to directly apply the method. Usually, in these cases a features selection procedure is needed before proceeding with quadratic analysis. Instead, by scoring the dataset with the Beta response function approach it is possible to perform the classification anyway as we reduce to operate on a single continuous variable. Moreover, the misclassification result obtained in this way is the same obtained with the other classification methods.

We do not present the comparison with the results obtained in the CoIL 2000 challenge. Indeed, in the competition only a sample of 5822 training units with known class labels has been provided to participants while our method has been tested considering a 10-fold cross validation on the whole set of 9822 observations. In addition, the task was only focused in predict which customers where interested in the caravan insurance policy and not in discriminating between who was interested or not. The results are therefore not directly comparable.

Table 6.2: *CoIL challenge 2000 dataset classification results.*

	Quantile	**LDA**	**QDA**	**SVM**	**NB**
Raw scores	0.1038 (0.0161)	0.0595 (0.0066)	-	0.0596 (0.0065)	0.1260 (0.0147)
Ridit scores	0.1009 (0.0125)	0.0594 (0.0067)	-	0.0596 (0.0065)	0.2873 (0.0273)
Normal median scores	0.1024 (0.0149)	0.0596 (0.0065)	-	0.0596 (0.0065)	0.1158 (0.0129)
Blom scores	0.1024 (0.0149)	0.0596 (0.0065)	-	0.0596 (0.0065)	0.1158 (0.0130)
NMS	0.1024 (0.0149)	0.0596 (0.0065)	-	0.0596 (0.0065)	0.1140 (0.0129)
LMS	0.1024 (0.0149)	0.0596 (0.0065)	-	0.0596 (0.0065)	0.1014 (0.0116)
LNMS	0.1028 (0.0148)	0.0596 (0.0065)	-	0.0596 (0.0065)	0.1291 (0.0152)
BRFA	0.0596 (0.0065)	0.0596 (0.0065)	0.0596 (0.0065)	0.0596 (0.0065)	0.0596 (0.0065)

Chapter 7

Conclusions and future research

7.1 Main findings of the study

The contribution of the present study is to propose a novel method for analysing and classifying data measured on an ordinal scale. Indeed, although a large amount of methods exists for dealing with situations where the response variables are measured on an ordinal scale, limited work has been done in the field of ordinal input features classification, despite these kind of data arise frequently in many different domains of research.

Ordinal data present unique challenges as they differ both from nominal and interval scaled data. They differ from the former in that they present order information and they differ from the latter as, if we follow the definition of ordinal scale introduced by Stevens *et al.* (1946), the concept of distance between categories is not present. For this reason we have dealt with the problem starting from the early stage of defining an appropriate metric distance between the ordinal categories.

We have defined a new methodology following several research steps.

First of all, we have done an exhaustive analysis of the state of art of the statistical literature on this topic with the aim of disentangling the most used approaches to ordinal data.

Three main ways to deal with ordered categorical variables emerged from this research: the parametric approach, which consists in replacing the categories with arbitrary numerical values and proceeding in the analysis by using the classical parametric inference methods; the nonparametric approach, which considers methods that only use ordering information about the categories, without making assumptions on the distribution of the ordinal variables and the underlying variable approach, which assigns numerical scores in such a way to meet as closely as possible some distributional assumption. In the last approach the scores should reflect the researcher's knowledge of an appropriate mathematical distance between the categories.

Considering the limitations and possible advantages of each of these three approaches, we have chosen to proceed with the underlying variable approach, which seemed the most adapt to our scope. In this framework the main interest, broadly speaking, is to evaluate whether it is possible to obtain acceptable classification results by assigning scores in a meaningful way.

From the literature review it emerged that several methods for assigning numeric values to categories, commonly known as scoring methods, have been developed over the years. We performed our simulations considering, among the others: the raw scores, the ridits, the blom scores, the normal median scores and the conditional mean scores (specifically, normal mean scores, logistic mean scores and log-normal mean scores).

The effectiveness of these methods, as far as we know, has never been tested for classification purposes.

The scoring systems present specific advantages as they have fast computational times and they provide a simple interpretation of the data. However, they also present some limitations as they do not allow to treat all the variables simultaneously, leading to a possible loss of information and, as seen in the case of log-normal mean scores, if the underlying distribution is misspecified this is likely to result in large classification errors.

In order to overcome the problems related to the scoring methods we have proposed, as possible solution, to operate inside the GLLVM framework, using the response function approach with a logit link function. This allows to treat the distribution of the whole ordinal response pattern, conditionally to a set of latent variables, without necessarily specifying a distributional form of the latent variables for each manifest variable.

Our method, like a generic scoring system, has the aim to shift from ordinal scaled to interval scaled input features in order to directly apply the standard classification methods.

In defining the most appropriate functional form for the latent variables we faced to the problem that the assumption of normally distributed latent variables is not always appropriate for classification tasks where data come from an unobserved heterogeneous populations.

Therefore, we have considered the work of Cagnone & Viroli (2014), where a finite mixture of multivariate Gaussians distributions is suggested for the latent variables in order to deal with data heterogeneity.

Although this approach undoubtedly presents advantages, as it allows for a flexible distributional form, it also has some limitations due to the fact that we loss in terms of computational efficiency as the number of components of the mixture increases.

The solution we propose is to still operate in the framework of response function approach, which seems to be a better alternative to the classical scoring methods, but choosing a different underlying distribution.

It seemed to us that the Beta latent distribution may be an appealing alternative for our scope, as it allows for a trade-off between flexibility and faster computational times. This distribution has never been applied in the latent variable models and, in particular, with the purpose of supervised classification of ordinal variables.

From this theoretical framework we have developed a dedicated algorithm with the aim of modelling the ordinal data and infer from the response patterns the associated values on a continuous scale. We have implemented an EM algorithm for parameters estimate. The integrals, which were not possible to solve analytically, have been approximated through Gauss-Legendre quadrature points.

Finally, we have compared the Beta response function approach with the scoring methods by performing a sensitivity analysis and by checking whether conclusions depend on the chosen set of scores.

We have compared the performance of five different classifiers (quantile-based, linear discriminant analysis, quadratic discriminant analysis, support vector machine and naive Bayes classifier) in terms of the mean misclassification rates from a 10-fold cross-validation.

In the simulation study, the data sets upon which the scoring methods have been applied have been generated according to four different scenarios in order to cover a set of possible situations. The scenarios include the cases of data sampled from symmetrical or highly skewed distributions. Moreover, the effect of specific factors such as the number of features, the sample size and the number of categories for the ordinal variables has been considered.

From the wide simulation study and from the real data example we presented, it emerges that the Beta response function approach can be a useful tool when we deal with ordinal variables and it allows to apply the standard classification methods directly. The properties of these classifiers on continuous data are known and they have been widely analysed.

Unfortunately, our proposal is not always the optimal one and sometimes alternative scoring methods should be preferred. However, in specific cases it may lead to a considerable gain in terms of correctly classified instances. It is therefore an opportune alternative to consider the BRFA in a validation set before proceeding with the classification.

We notice that the classification results obtained with BRFA are particularly good considering that they are based on a single latent variable. As mentioned, a possible explanation is that the dependence structure of the data is taken into account because the scores are obtained conditionally to the p-dimensional response pattern. In the CoiL challenge dataset, where the features are highly correlated, the Beta response function approach resulted in the best classification performance, independently from the chosen classifier. As shown in the following section, it is our scope to extend the method also to the case of more than one latent variable.

From the graphs in the appendix it is possible to notice that the classification results associated with the scoring methods present similar patterns among datasets where different number of categories have been considered.

Differences in the results associated with raw scores, ridits score, blom scores, and conditional mean scores are generally not significant. There is no method that is uniformly superior to the others. The differences between category scores attributed with these scoring systems do not appear to be large enough to result in significant variation in the misclassification rates. A separate case is given by the log-normal mean scores, which, as mentioned in chapter 5, are particularly unsuitable in almost all the considered scenarios with the exception of scenario 3, where data are generated from a negatively skewed continuous underlying distributions. For scenario 3, the misclassification rates obtained through LNMS greatly improve and, in some cases, outperforms the BRFA results.

In particular, from the simulations emerge that, whenever the dataset is generated from populations whose distributional form is symmetrical (specifically in this simulation study the Student's t distribution) for quantile-based, LDA and QDA the results associated with the BRFA are almost always significantly better than those of all other scoring methods.

The case where the contribution of our method is more evident is that of quadratic discriminant analysis. Indeed, this classifier is particularly unsuitable for this type of data and the results obtained by directly applying it to raw scores (as well as the other scoring methods introduced in chapter 2) are considerably worse with respect to all other classifiers.

Instead, applying the classifier to data scored through the BRFA, as it operates on a single continuous variable, the results significantly improve and they become comparable to those obtained through the other classification methods.

When the observed variables are collinear, as in the real data example presented in the previous chapter, it is known that quadratic discriminant analysis can not be directly applied since it is not possible to obtain parameter estimates with non-invertible covariance matrices. For this reason, sometimes a feature selection procedure is required in order to apply the classifier. In certain circumstances, our method can be seen as an alternative to feature selection. It may also lead to a significant improvement in the classification results.

A secondary but nonetheless important aspect of BRFA is that, because it operates in the framework of factorial analysis, it also allows to perform a graphical analysis of the available data, which are synthesized into a single continuous variable. This is often a useful tool since, as they say, “an image is worth more than a thousand words”.

The Beta response function approach resulted to not be the optimal method for assigning scores in the simulation study when SVM or naive Bayes classifier are considered (especially in scenario 2 and 3 where data are highly skewed and well separated). These classifiers generally provide for the best results in any simulated dataset. Anyhow, when the number of features is high (i.e. $p = 40, 100$), the misclassification rates obtained through any scoring method are close to 0 and do not significantly differ from each other.

The fact that the BRFA does not work properly with SVM and naive Bayes in the simulations should not prevent the use of our method with these classifiers. As a matter of fact, in the large macrocosm of classification methods, there is not a classifier that is uniformly superior to the others in every circumstance. Indeed, in the CoIL dataset the naive Bayes classifier do not provide for the best performance but the misclassification rate reduces to about the half if it is applied on BRFA data.

7.2 Ongoing work

As shown in chapter 3, our proposal leads to shorter computational times than those observed considering a finite mixture of Gaussians distributions (see Table 3.3). However, the computational times are still quite high for an easy practical application.

The following table shows the computational times (in minutes) obtained by applying the Beta response function approach to features simulated by discretizing from a multivariate Gaussian distribution. Computational times are recorded for different combinations of sample size (N) and number of features (p).

The times displayed in the table are the averages over 10 running of the algorithm with different initial seeds.

Table 7.1: *Computational times (in minutes) required for computing scores with BRFA.*

	$\boldsymbol{p = 10}$	$\boldsymbol{p = 20}$	$\boldsymbol{p = 40}$	$\boldsymbol{p = 100}$
$\boldsymbol{N = 100}$	00:52	01:36	03:58	09:39
$\boldsymbol{N = 200}$	01:58	03:52	08:36	20:13
$\boldsymbol{N = 400}$	03:47	08:19	14:53	40:16

As it is possible to notice from Table 7.1, the computational times are very high, especially for an increasing number of features.

We are currently trying to speed up the algorithm. The possible solutions we have considered so far are two:

1. Consider alternative numerical optimization algorithms respect to those selected for parameters estimation in the EM algorithm (see chapter 3). Indeed, these algorithms are the main cause of the code slowdown.

2. Consider an intermediate step of the **R** code in **C**, which is a general-purpose computer programming language.

Once the code will has been speeded up, the next step will be to consider a larger number of latent variables in the model.

A further application of BRFA can also be found in clustering, where the concept of distance between observations is often used to define groups of similar instances. We expect possible good results in applying our method also in the field of cluster analysis. This can be a further guideline for future researches.

Appendices

Appendix A

Other simulation results

The appendix is organized as follows: each section presents the simulation results, respectively for: quantile-based, linear discriminant analysis, quadratic discriminant analysis, support vector machine and naive Bayes classifier. For each classifier, plots for the mean misclassification rates (for every combination of p and N) in the four considered scenarios are reported, along with a summary table that shows the misclassification rates and the related standard errors.

The sections are further subdivided on the base of the number of categories considered for the features in the data set (i.e. $C = 3, 5, 357$).

With $C = 357$ we identify the cases where features with different number of categories (i.e. 3, 5 and 7 categories) are considered inside the same data set. In these cases we consider $p = 15, 30, 60$ so that features with the same number of categories in the data set can be equally subdivided.

The plots have the same structure of the ones presented in chapter 5, i.e. each block represent a different number of features. Inside each block the mean misclassification rates from the 10-fold cross-validation are reported. Black symbols are used for the mean misclassification rates obtained by applying the classifier on the scoring methods from chapter 2 while the red dots are for the BRFA results.

A.1 Quantile-based classifier

Similarly to what observed in chapter 5, in the case $C = 3$ (figures A.1 to A.4) it is possible to notice that in scenario 1 there is a clear advantage in applying the quantile-based classifier on BRFA data. In scenarios 2, 3 and "mixture" we have that, when the data set dimension is relatively small (i.e. $p = 10, 20$), the BRFA does not seem to be the optimal approach as the results are worse with respect to any other scoring method. Instead, in high dimensional settings we have that no significant differences in the classifier performance among the scoring methods are observed in the majority of situations. The exceptions are the cases where the number of features is high ($p = 40, 100$) and the number of instances is small ($N = 100$). In particular in scenario 3 (for $p = 100$ and $N = 100$) and "mixture" scenario (for $p = 40, 100$ and $N = 100$) the BRFA data set provides better results.

Figures A.5 to A.8 report classification results for simulated datasets composed by features with different number of categories. In the first scenario we observe that, as for simulations where the features have all the same number of categories, the quantile-based classifier performs at its best on data scored through BRFA. In all other scenarios the BRFA does not provide the best classification result for every combination of N and p. However, when p is high the misclassification rate is comparable to the ones obtained through ridits and conditional mean scores.

A.1.1 Simulation results for $C = 3$

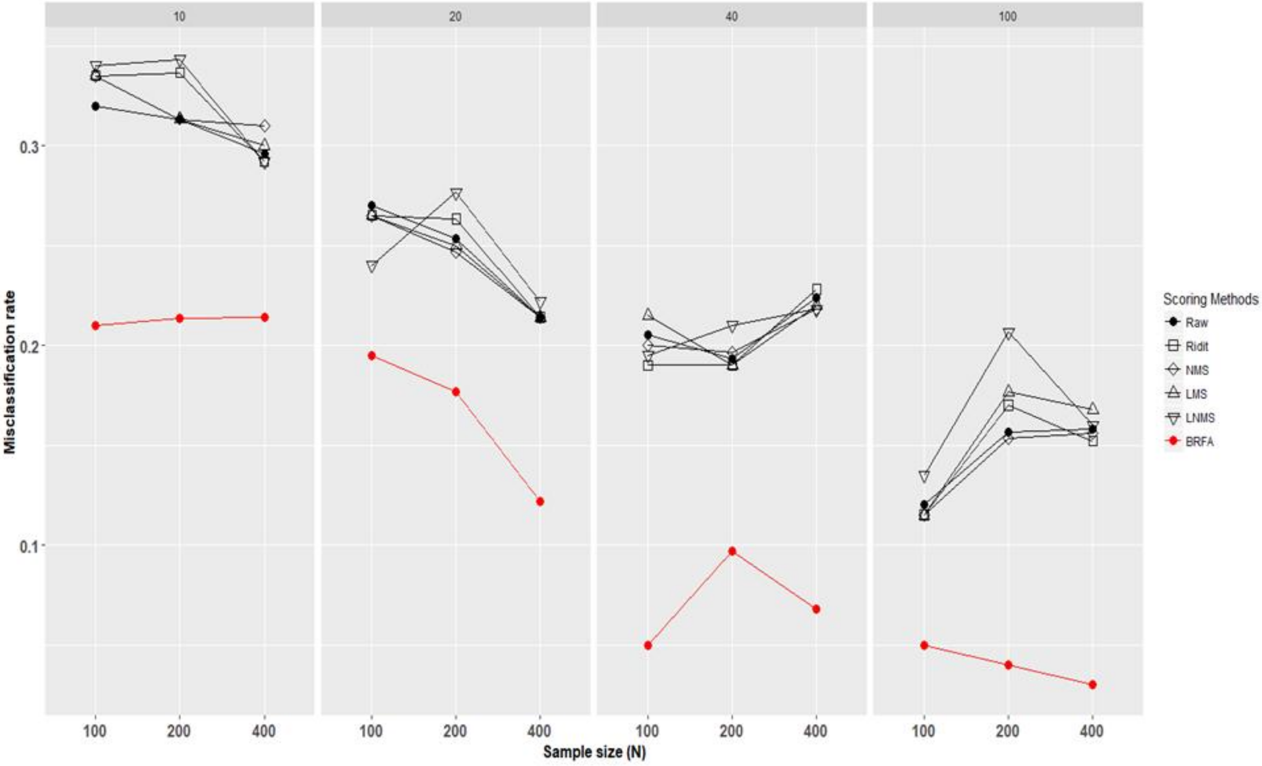

Figure A.1: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

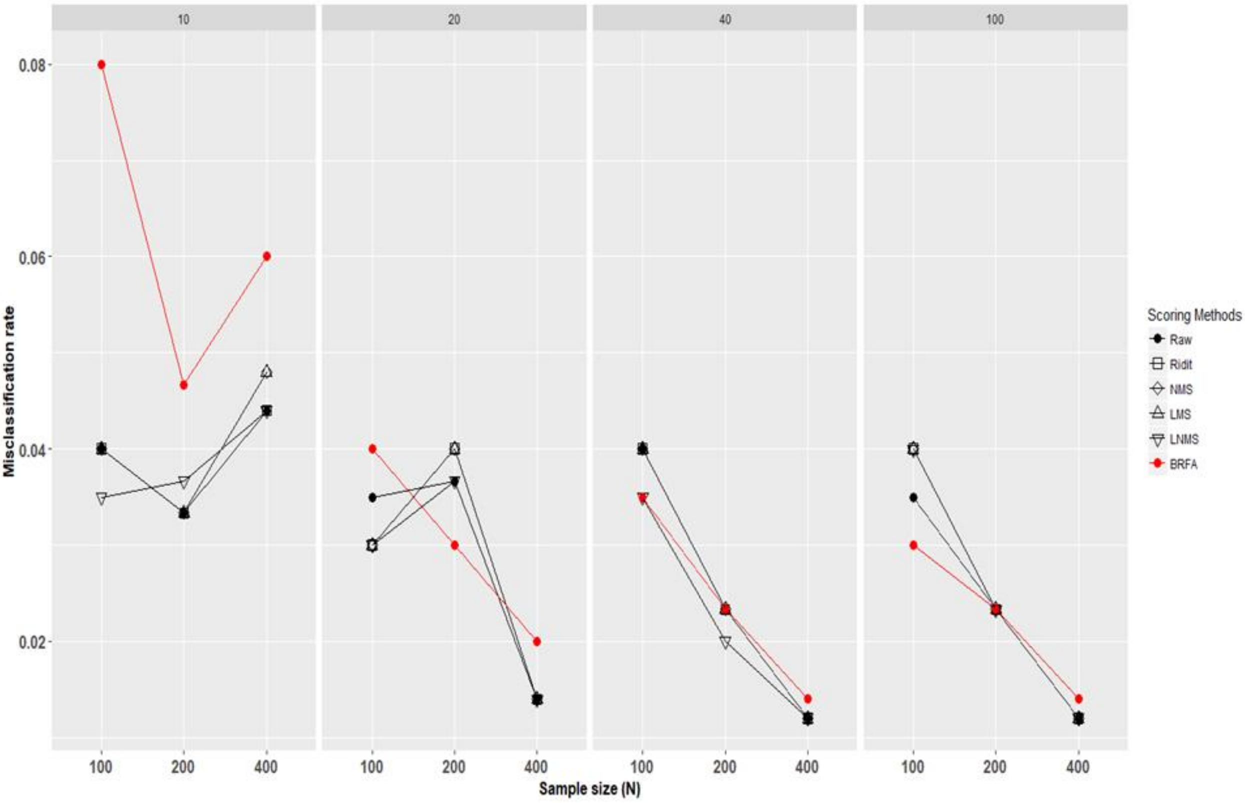

Figure A.2: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 3$ *categories.*

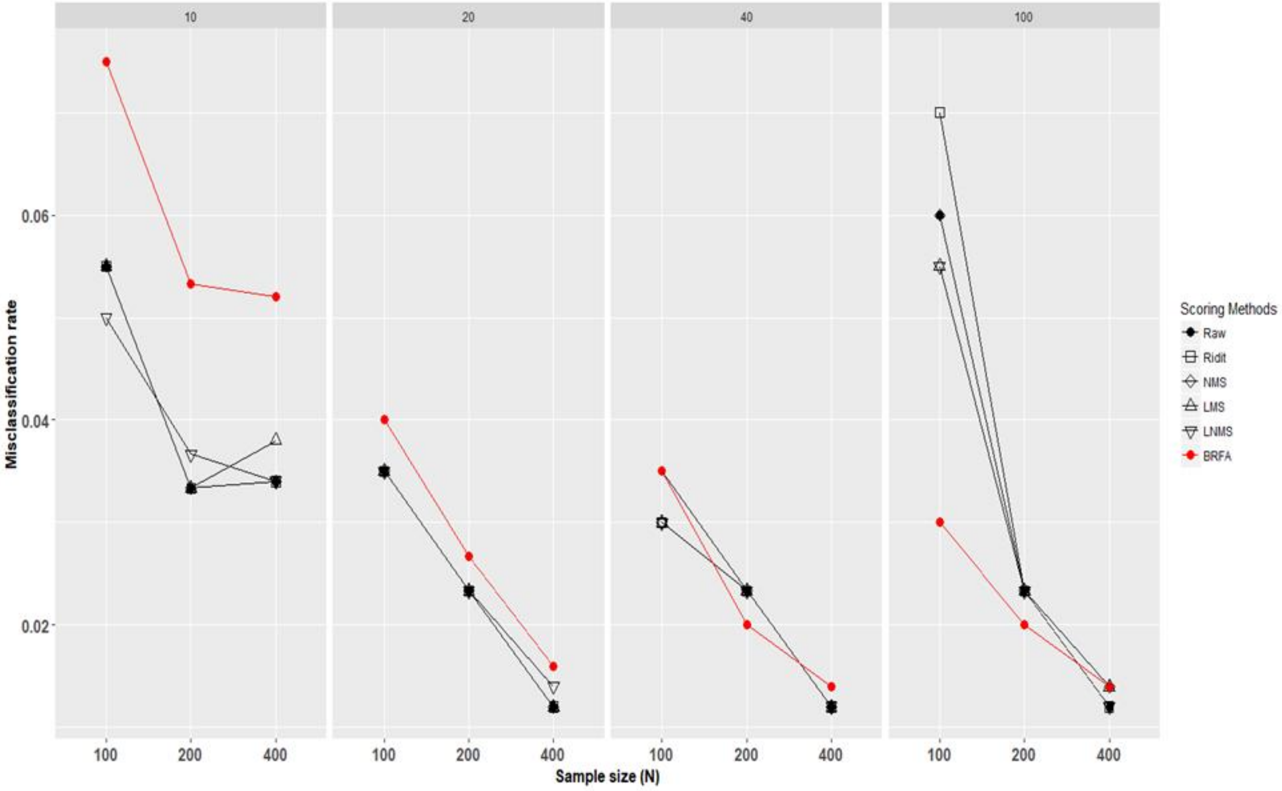

Figure A.3: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 3$ *categories.*

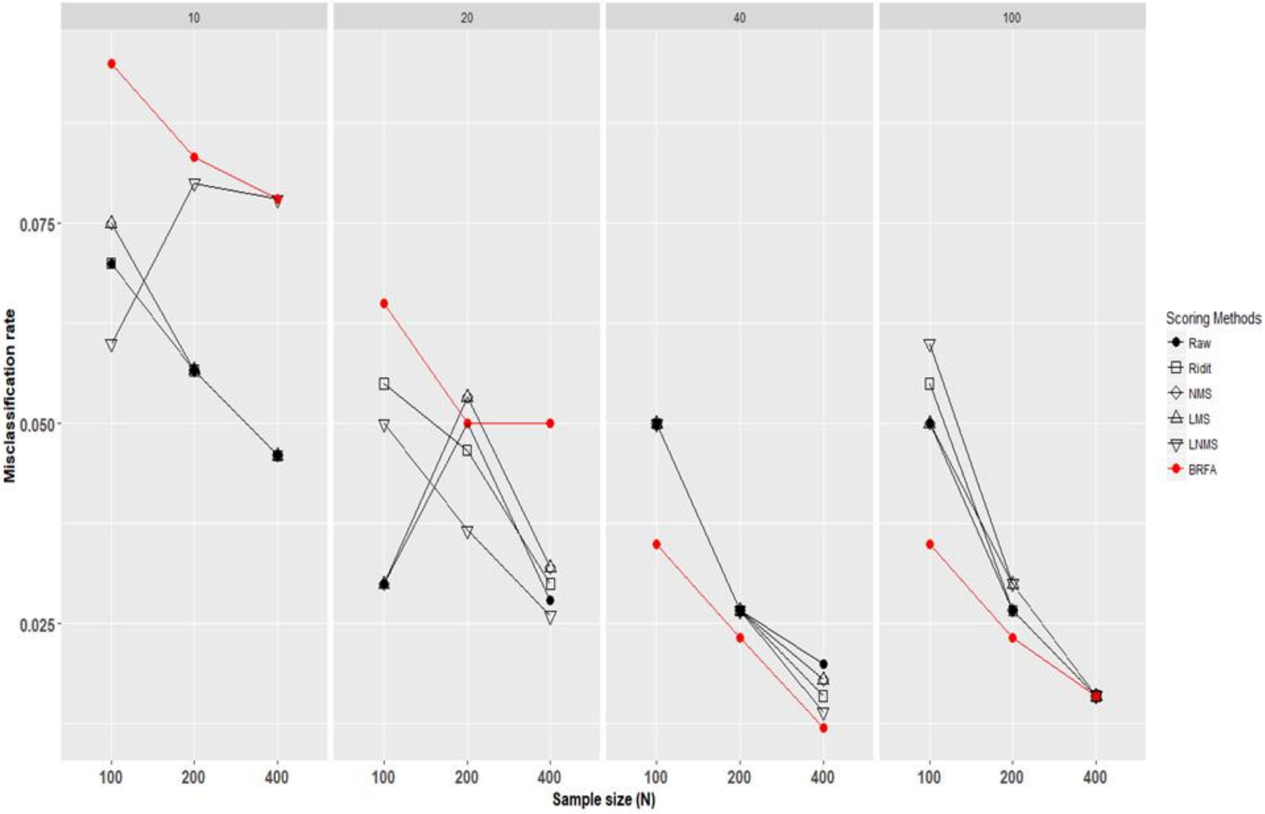

Figure A.4: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

Table A.1: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for quantile-based classifier in the four considered scenarios. Simulation results refer to the case of ordinal features with three categories (C = 3).*

						Scenario 1						
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.32 (0.03)	0.31 (0.02)	0.30 (0.02)	0.27 (0.03)	0.25 (0.03)	0.21 (0.02)	0.20 (0.03)	0.19 (0.01)	0.22 (0.02)	0.12 (0.02)	0.16 (0.02)	0.16 (0.02)
Ridit	0.34 (0.02)	0.34 (0.01)	0.29 (0.02)	0.26 (0.03)	0.26 (0.03)	0.21 (0.02)	0.19 (0.03)	0.19 (0.02)	0.23 (0.02)	0.12 (0.02)	0.17 (0.02)	0.15 (0.02)
NM	0.34 (0.03)	0.31 (0.02)	0.30 (0.02)	0.26 (0.03)	0.25 (0.03)	0.21 (0.02)	0.22 (0.03)	0.19 (0.01)	0.22 (0.02)	0.12 (0.03)	0.18 (0.02)	0.17 (0.01)
Blom	0.34 (0.03)	0.31 (0.02)	0.30 (0.02)	0.26 (0.03)	0.25 (0.03)	0.21 (0.02)	0.22 (0.03)	0.19 (0.01)	0.22 (0.02)	0.12 (0.03)	0.18 (0.02)	0.17 (0.01)
NMS	0.34 (0.02)	0.31 (0.02)	0.31 (0.02)	0.26 (0.03)	0.25 (0.03)	0.21 (0.02)	0.20 (0.03)	0.2 (0.01)	0.22 (0.01)	0.12 (0.03)	0.15 (0.02)	0.16 (0.02)
LMS	0.34 (0.03)	0.31 (0.02)	0.3 (0.02)	0.26 (0.03)	0.25 (0.03)	0.21 (0.02)	0.22 (0.03)	0.19 (0.01)	0.22 (0.02)	0.12 (0.03)	0.18 (0.02)	0.17 (0.01)
LNMS	0.34 (0.02)	0.34 (0.02)	0.29 (0.02)	0.24 (0.02)	0.28 (0.03)	0.22 (0.02)	0.20 (0.03)	0.21 (0.02)	0.22 (0.02)	0.14 (0.03)	0.21 (0.01)	0.16 (0.02)
BRFA	0.21 (0.03)	0.21 (0.02)	0.21 (0.01)	0.20 (0.03)	0.18 (0.01)	0.12 (0.01)	0.05 (0.02)	0.10 (0.02)	0.07 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)
						Scenario 2						
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.04 (0.02)	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)
Ridit	0.04 (0.02)	0.03 (0.01)	0.04 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
NM	0.04 (0.02)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
Blom	0.04 (0.02)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
NMS	0.04 (0.02)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
LMS	0.04 (0.02)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
LNMS	0.04 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.04 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)
BRFA	0.08 (0.02)	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
						Scenario 3						
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.07 (0.02)	0.02 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.03 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
LNMS	0.05 (0.02)	0.04 (0.01)	0.03 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
BRFA	0.08 (0.01)	0.05 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
						Mixture scenario						
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Ridit	0.07 (0.02)	0.06 (0.01)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.06 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Blom	0.08 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NMS	0.08 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
LMS	0.08 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
LNMS	0.06 (0.02)	0.08 (0.02)	0.08 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.01 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.1 (0.02)	0.08 (0.02)	0.08 (0.01)	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)

A.1.2 Simulation results for $C = 357$

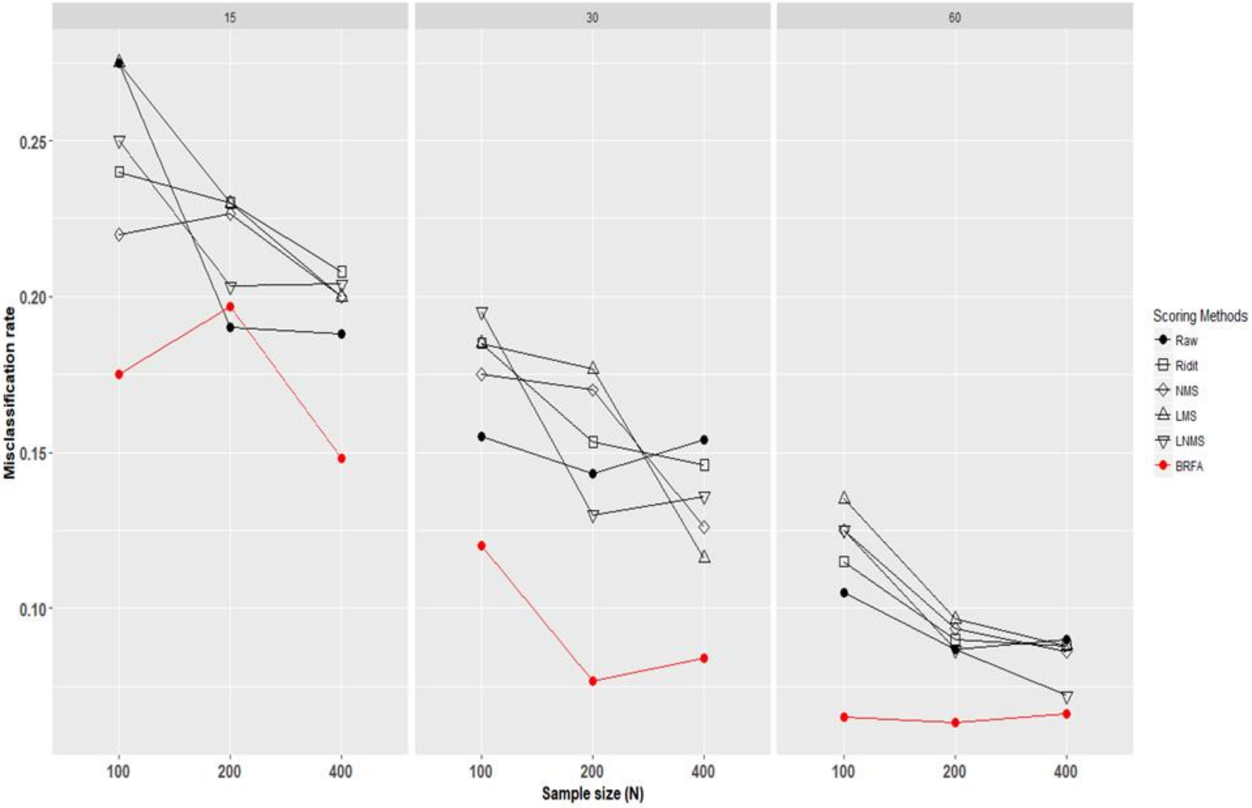

Figure A.5: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories $C = 357$.*

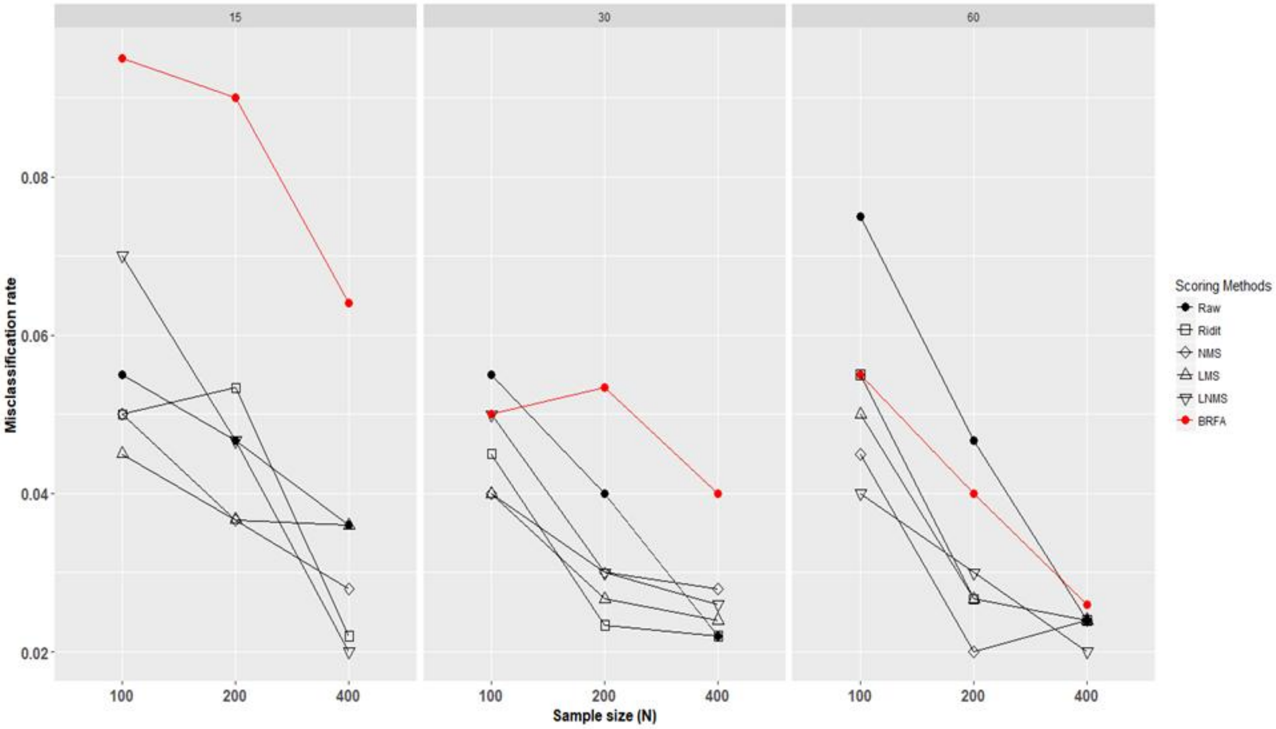

Figure A.6: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

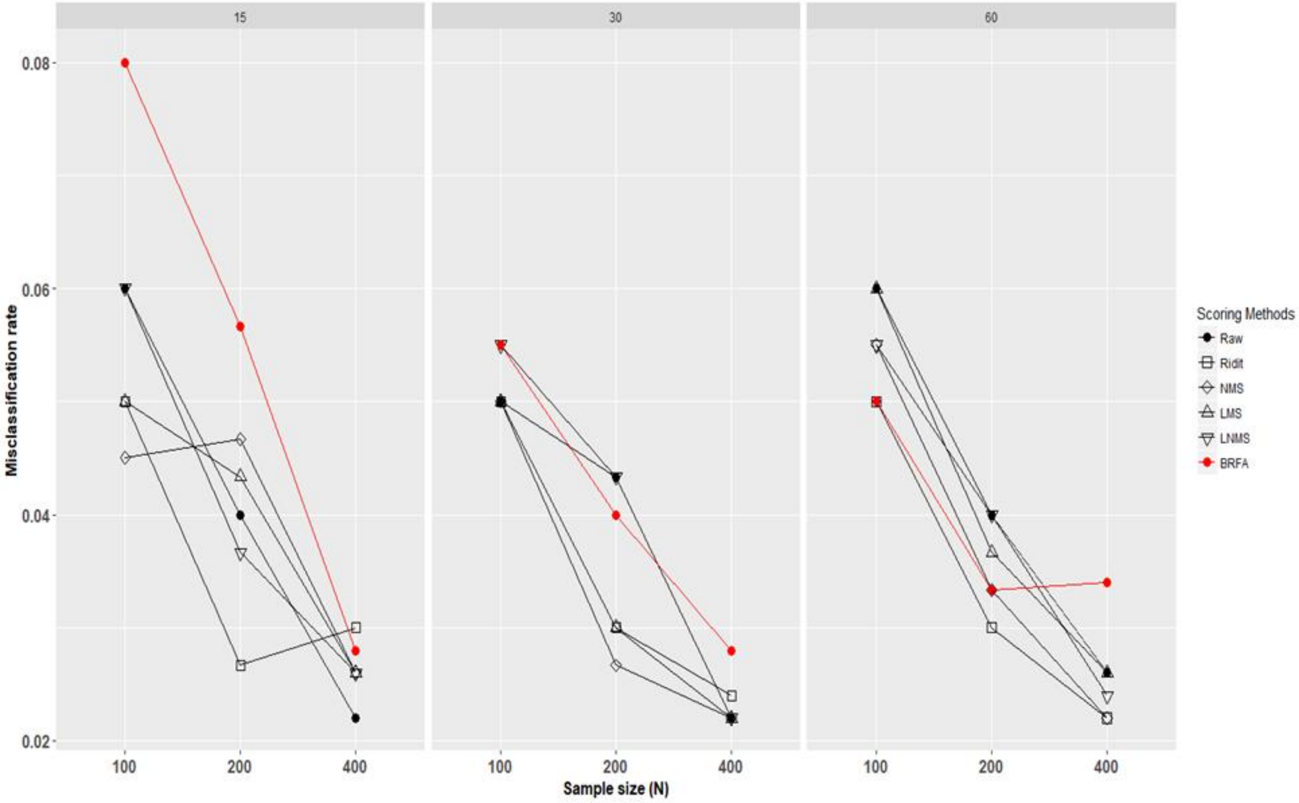

Figure A.7: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

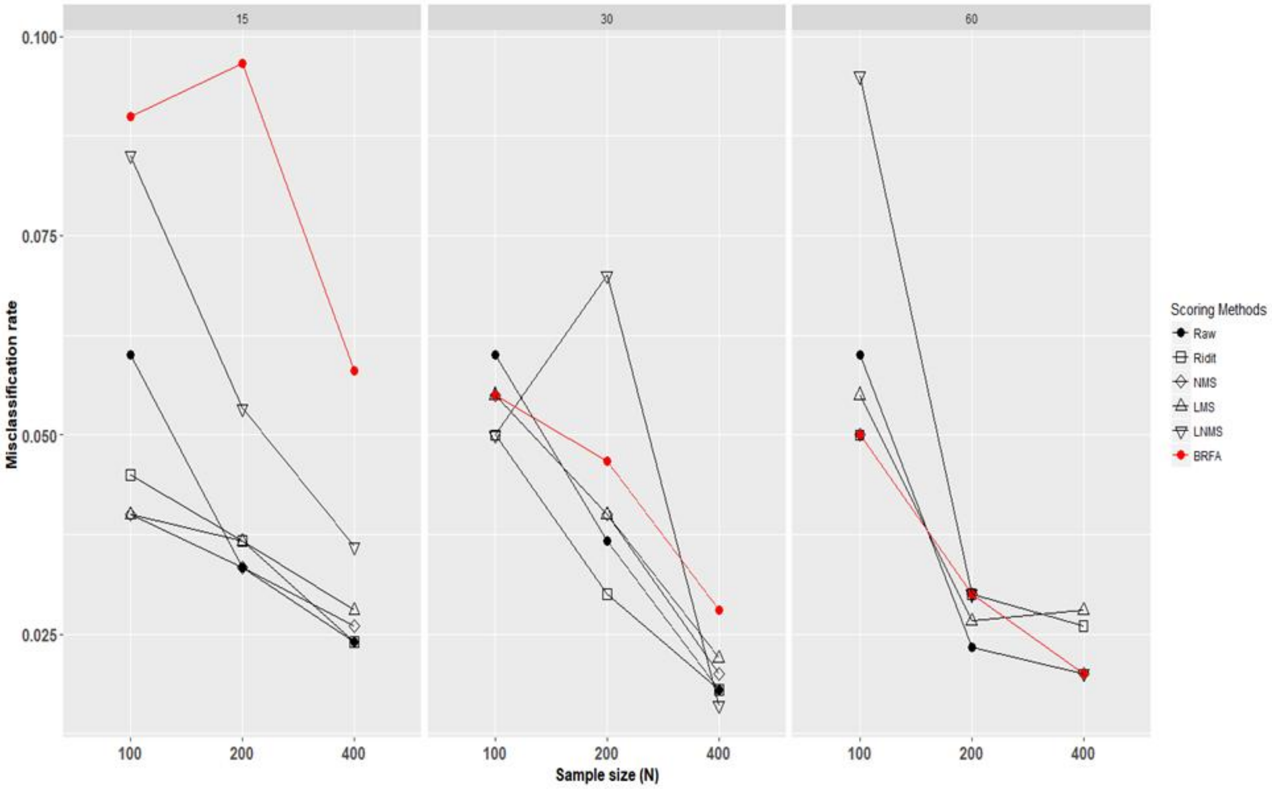

Figure A.8: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

Table A.2: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for quantile-based classifier in the four considered scenarios. Simulation results refer to the case of ordinal features with different number of categories ($C = 357$).*

	Scenario 1								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.28 (0.03)	0.19 (0.02)	0.19 (0.02)	0.16 (0.03)	0.14 (0.03)	0.15 (0.02)	0.10 (0.03)	0.09 (0.02)	0.09 (0.01)
Ridit	0.24 (0.04)	0.23 (0.03)	0.21 (0.02)	0.18 (0.03)	0.15 (0.02)	0.15 (0.02)	0.12 (0.02)	0.09 (0.02)	0.09 (0.01)
NM	0.21 (0.03)	0.23 (0.03)	0.20 (0.02)	0.17 (0.03)	0.16 (0.03)	0.12 (0.02)	0.12 (0.02)	0.09 (0.01)	0.09 (0.01)
Blom	0.21 (0.03)	0.23 (0.03)	0.20 (0.02)	0.17 (0.03)	0.16 (0.03)	0.12 (0.02)	0.12 (0.02)	0.09 (0.01)	0.09 (0.01)
NMS	0.22 (0.04)	0.23 (0.03)	0.20 (0.02)	0.18 (0.02)	0.17 (0.03)	0.13 (0.01)	0.12 (0.02)	0.09 (0.02)	0.09 (0.01)
LMS	0.28 (0.03)	0.23 (0.03)	0.20 (0.02)	0.18 (0.03)	0.18 (0.03)	0.12 (0.01)	0.14 (0.02)	0.10 (0.02)	0.09 (0.01)
LNMS	0.25 (0.04)	0.20 (0.03)	0.20 (0.01)	0.20 (0.02)	0.13 (0.03)	0.14 (0.02)	0.12 (0.02)	0.09 (0.02)	0.07 (0.01)
BRFA	0.18 (0.03)	0.20 (0.01)	0.15 (0.02)	0.12 (0.02)	0.08 (0.01)	0.08 (0.01)	0.06 (0.02)	0.06 (0.01)	0.07 (0.01)
	Scenario 2								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
Ridit	0.05 (0.01)	0.05 (0.01)	0.02 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
NM	0.05 (0.01)	0.06 (0.01)	0.03 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Blom	0.05 (0.01)	0.06 (0.01)	0.03 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NMS	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.04 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)
LMS	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
LNMS	0.07 (0.02)	0.05 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.10 (0.03)	0.09 (0.02)	0.06 (0.01)	0.05 (0.01)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
	Scenario 3								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
Ridit	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
LMS	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
LNMS	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
BRFA	0.08 (0.01)	0.06 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)
	Mixture scenario								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.02)	0.02 (0)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)
Ridit	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.03 (0.01)
NM	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
Blom	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
NMS	0.04 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LMS	0.04 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)
LNMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.07 (0.02)	0.02 (0.01)	0.10 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.09 (0.02)	0.10 (0.02)	0.06 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)

A.2 Linear discriminant analysis

In this section, as well as for the QDA, SVM and Naive Bayes classifier sections, we include the results for the case $C = 5$ that have not been presented in chapter 5. We notice that, as already mentioned, the LDA has worse performance than the other classifiers (quantile-based, SVM and Naive Bayes), independently of the scenario and the scoring method adopted. From figure A.9 to A.11, in accordance with what observed in the previous chapters, it is possible to notice that in scenario 2 and "mixture" scenario the LNMS are particularly unsuitable for this classifier. Referring to scenario 3, when the number of instances is small and the number of features is high ($N = 200$ and $p = 100$) there is a clear advantage in using the BRFA. In general, when the number of features is high, the results obtained with BRFA are always comparable with the ones obtained with any other considered scoring method.
Also in the case $C = 3$ and $C = 357$ (figures A.12 to A.19) there is a gain in applying the LDA on BRFA in high dimensional settings in scenario 1. Unfortunately, in scenario 2 and 'mixture" scenario when the number of features is small the performance associated to the BRFA are generally worse. In these scenarios the results obtained through BRFA do not significantly differ from the other scoring methods when p is high. For $C = 3$ we have that in scenario 3 there are not significant differences between BRFA and the other scoring methods with the exception of the case $N = 100$ and $p = 40$ and the case $N = 200$ and $p = 100$, where the BRFA is the best method, together with LNMS.

A.2.1 Simulation results for $C = 5$

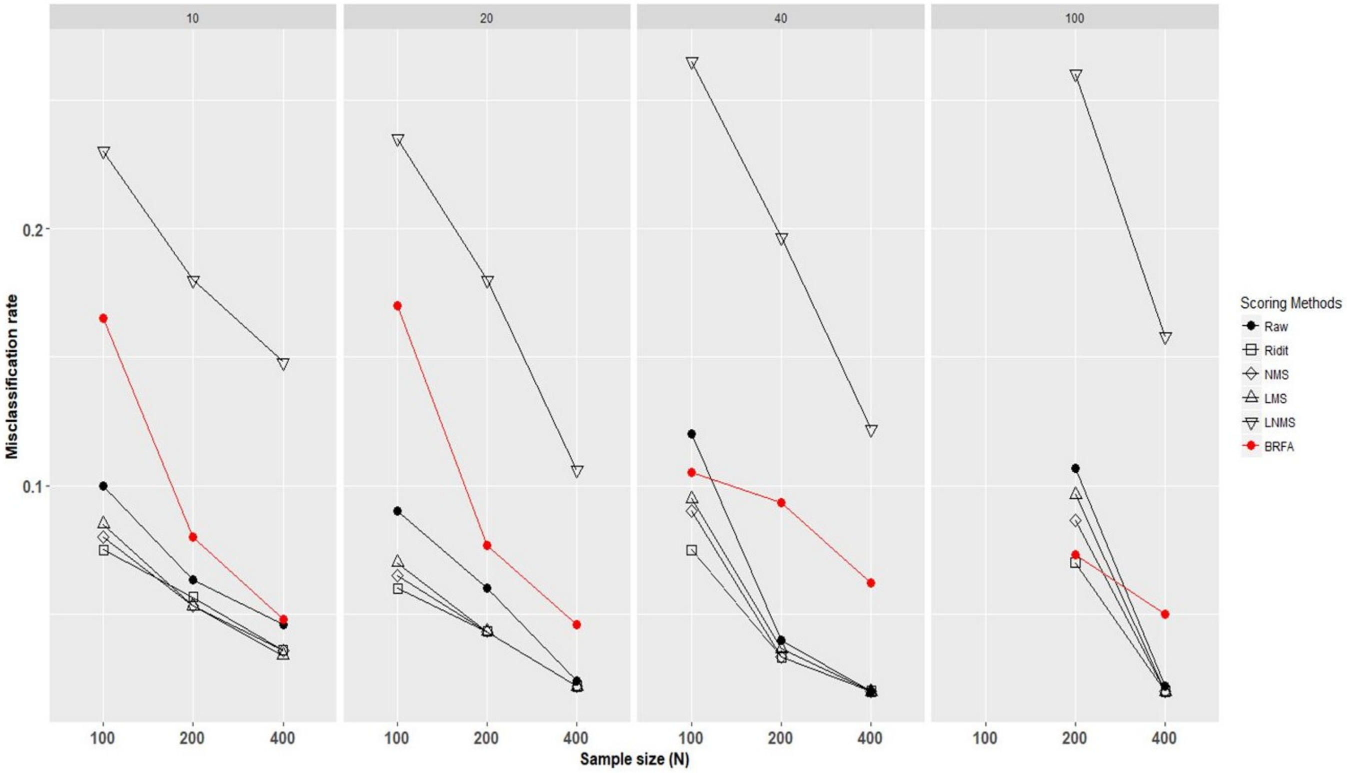

Figure A.9: *Results for Scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

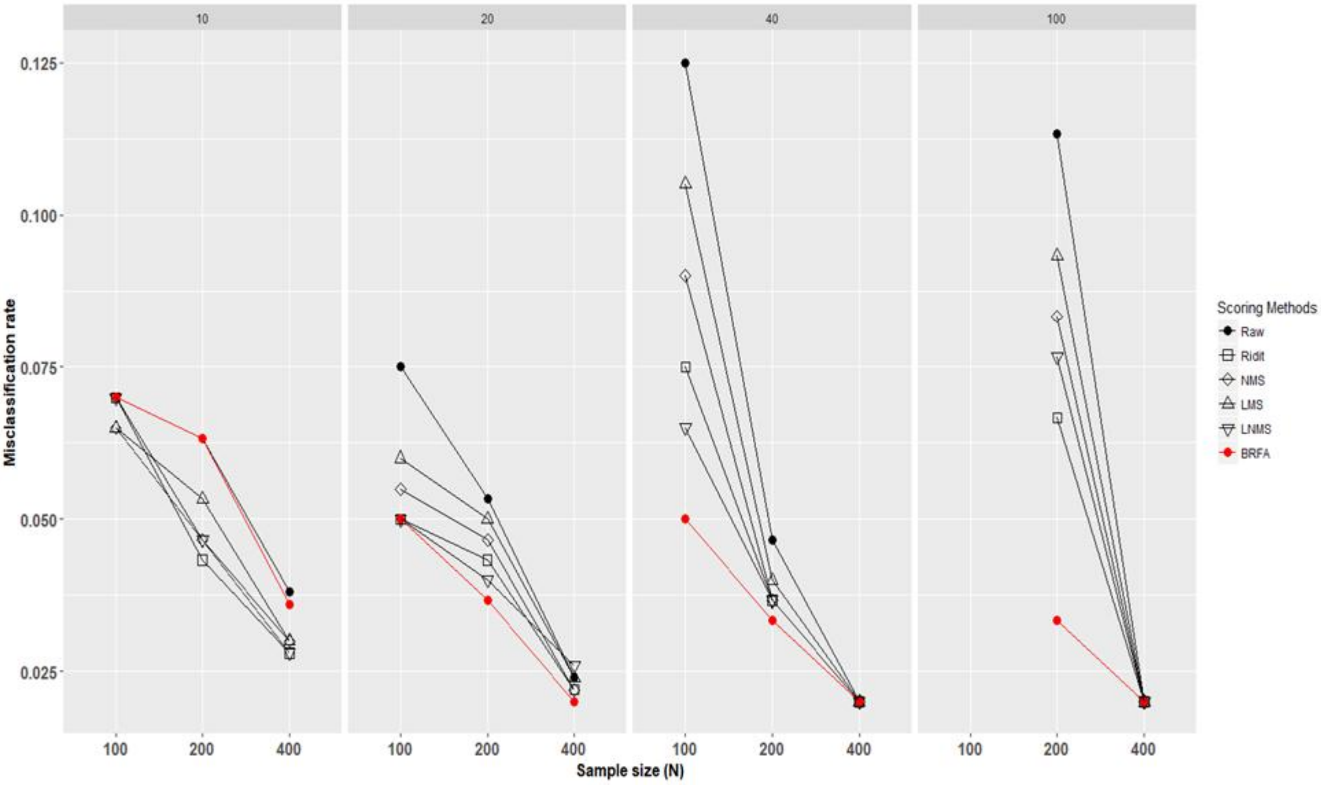

Figure A.10: *Results for Scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

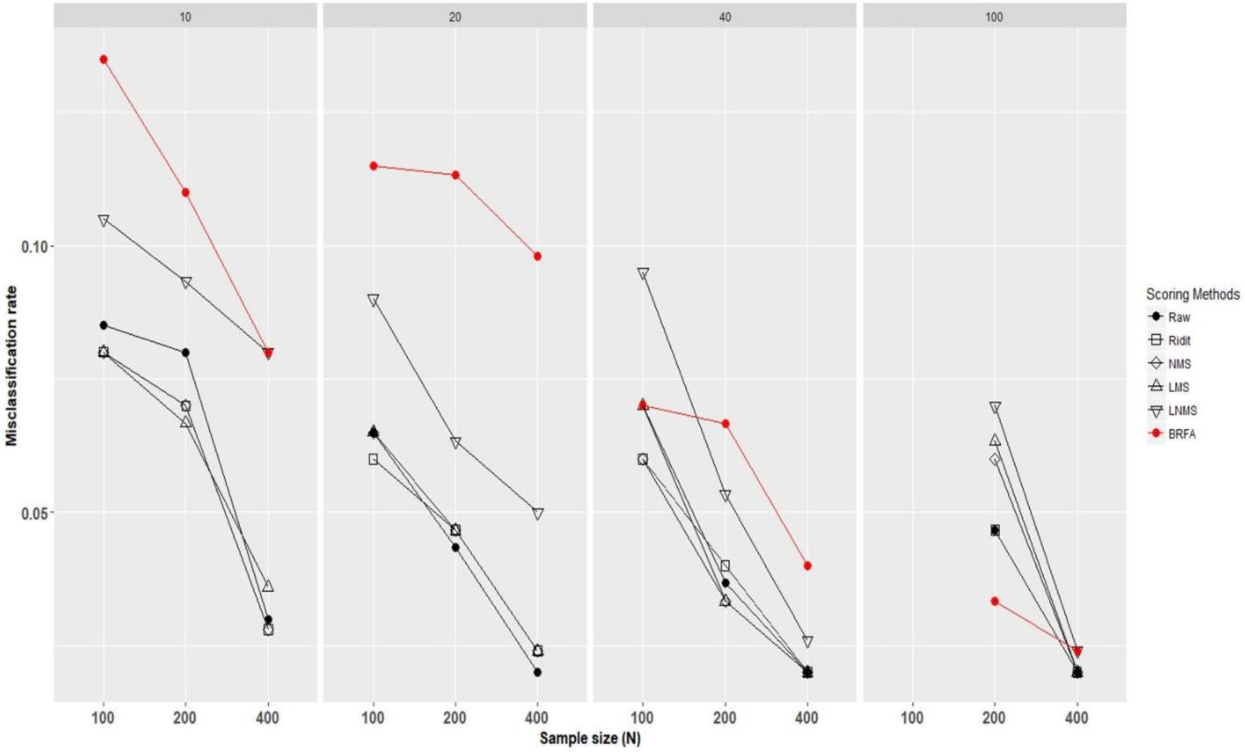

Figure A.11: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

Table A.3: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for LDA in the four considered scenarios. Simulation results refer to the case of ordinal features with five categories ($C = 5$).*

	Scenario 1											
	p=10			p=20			p=40			p=100		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.27 (0.02)	0.21 (0.02)	0.18 (0.02)	0.29 (0.03)	0.22 (0.03)	0.12 (0.01)	0.29 (0.02)	0.18 (0.03)	0.10 (0.01)	-	0.34 (0.03)	0.15 (0.02)
Ridit	0.23 (0.03)	0.20 (0.02)	0.18 (0.02)	0.26 (0.02)	0.16 (0.02)	0.12 (0.02)	0.23 (0.03)	0.14 (0.03)	0.08 (0.01)	-	0.31 (0.03)	0.09 (0.01)
NM	0.25 (0.04)	0.21 (0.02)	0.19 (0.02)	0.30 (0.03)	0.20 (0.03)	0.12 (0.01)	0.26 (0.02)	0.17 (0.02)	0.09 (0.01)	-	0.34 (0.02)	0.14 (0.01)
Blom	0.25 (0.04)	0.21 (0.02)	0.19 (0.02)	0.30 (0.03)	0.20 (0.03)	0.12 (0.01)	0.26 (0.02)	0.17 (0.02)	0.09 (0.01)	-	0.34 (0.02)	0.14 (0.01)
NMS	0.24 (0.03)	0.21 (0.02)	0.19 (0.02)	0.30 (0.03)	0.20 (0.03)	0.12 (0.01)	0.26 (0.03)	0.17 (0.02)	0.10 (0.01)	-	0.34 (0.03)	0.14 (0.01)
LMS	0.26 (0.04)	0.22 (0.02)	0.19 (0.02)	0.31 (0.03)	0.22 (0.03)	0.12 (0.01)	0.30 (0.03)	0.18 (0.02)	0.11 (0.01)	-	0.38 (0.03)	0.15 (0.02)
LNMS	0.34 (0.03)	0.34 (0.03)	0.32 (0.02)	0.34 (0.03)	0.32 (0.03)	0.29 (0.02)	0.27 (0.02)	0.30 (0.03)	0.26 (0.02)	-	0.35 (0.02)	0.26 (0.02)
BRFA	0.22 (0.02)	0.22 (0.03)	0.19 (0.02)	0.20 (0.03)	0.20 (0.03)	0.15 (0.01)	0.06 (0.01)	0.08 (0.02)	0.07 (0.01)	-	0.05 (0.01)	0.04 (0.01)
	Scenario 2											
	p=10			p=20			p=40			p=100		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.10 (0.02)	0.06 (0.01)	0.05 (0.01)	0.09 (0.02)	0.06 (0.01)	0.02 (0.01)	0.12 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.11 (0.01)	0.02 (0.01)
Ridit	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.08 (0.02)	0.03 (0.01)	0.02 (0.01)	-	0.07 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.06 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.09 (0.02)	0.03 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
Blom	0.08 (0.02)	0.06 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.09 (0.02)	0.03 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
NMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.09 (0.02)	0.03 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
LMS	0.08 (0.01)	0.05 (0.01)	0.03 (0.01)	0.07 (0.01)	0.04 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.10 (0.02)	0.02 (0.01)
LNMS	0.23 (0.04)	0.18 (0.02)	0.15 (0.01)	0.24 (0.03)	0.18 (0.02)	0.11 (0.01)	0.26 (0.02)	0.20 (0.02)	0.12 (0.02)	-	0.26 (0.02)	0.16 (0.01)
BRFA	0.16 (0.03)	0.08 (0.01)	0.05 (0.01)	0.17 (0.03)	0.08 (0.01)	0.05 (0.01)	0.1 (0.02)	0.09 (0.01)	0.06 (0.01)	-	0.07 (0.01)	0.05 (0.01)
	Scenario 3											
	p=10			p=20			p=40			p=100		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)	0.12 (0.03)	0.05 (0.01)	0.02 (0.01)	-	0.11 (0.02)	0.02 (0.01)
Ridit	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.08 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.07 (0.01)	0.02 (0.01)
NM	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.08 (0.03)	0.04 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
Blom	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.08 (0.03)	0.04 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.09 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.08 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.01)	0.05 (0.01)	0.02 (0.01)	0.10 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.09 (0.01)	0.02 (0.01)
LNMS	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.08 (0.01)	0.02 (0.01)
BRFA	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.02 (0.01)
	Mixture scenario											
	p=10			p=20			p=40			p=100		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.08 (0.02)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.01)	0.04 (0.01)	0.02 (0.01)	-	0.05 (0.02)	0.02 (0.01)
Ridit	0.08 (0.02)	0.07 (0.02)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	-	0.05 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.07 (0.02)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.06 (0.02)	0.02 (0.01)
Blom	0.08 (0.02)	0.07 (0.02)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.06 (0.02)	0.02 (0.01)
NMS	0.08 (0.02)	0.07 (0.02)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.06 (0.02)	0.02 (0.01)
LMS	0.08 (0.02)	0.07 (0.01)	0.04 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.07 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.06 (0.01)	0.02 (0.01)
LNMS	0.10 (0.02)	0.09 (0.01)	0.08 (0.01)	0.09 (0.02)	0.06 (0.01)	0.05 (0.01)	0.10 (0.02)	0.05 (0.01)	0.03 (0.01)	-	0.07 (0.02)	0.02 (0.01)
BRFA	0.14 (0.02)	0.11 (0.02)	0.08 (0.01)	0.12 (0.03)	0.11 (0.02)	0.10 (0.01)	0.07 (0.02)	0.07 (0.01)	0.04 (0.01)	-	0.03 (0.01)	0.02 (0.01)

A.2.2 Simulation results for $C = 3$

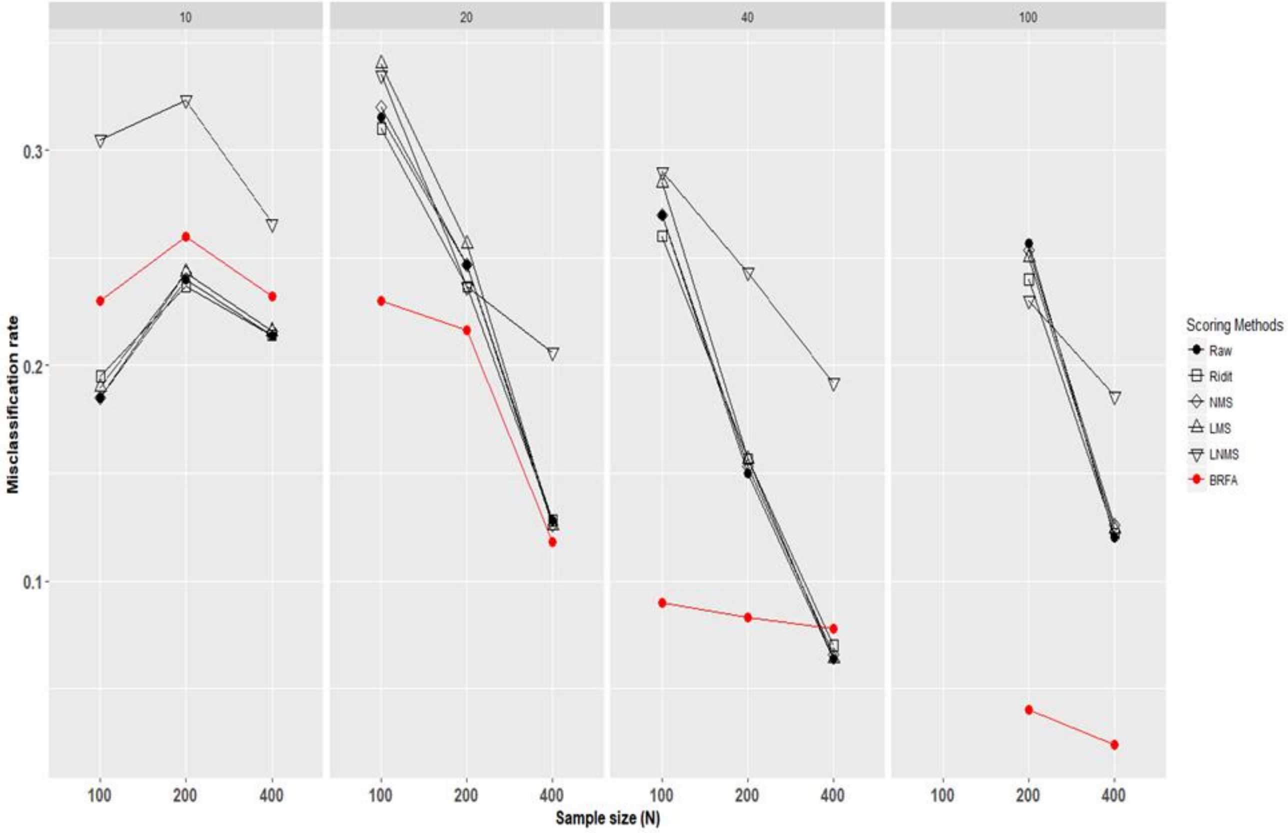

Figure A.12: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

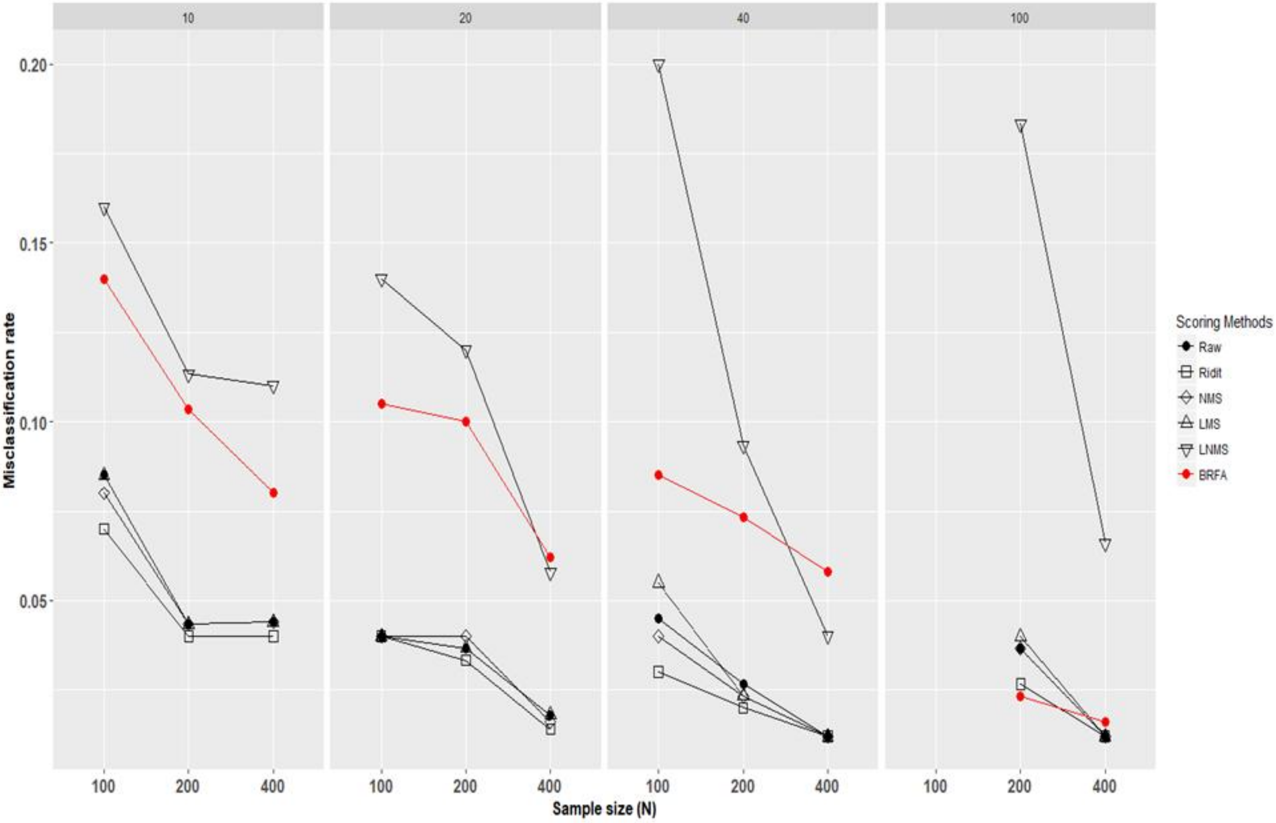

Figure A.13: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

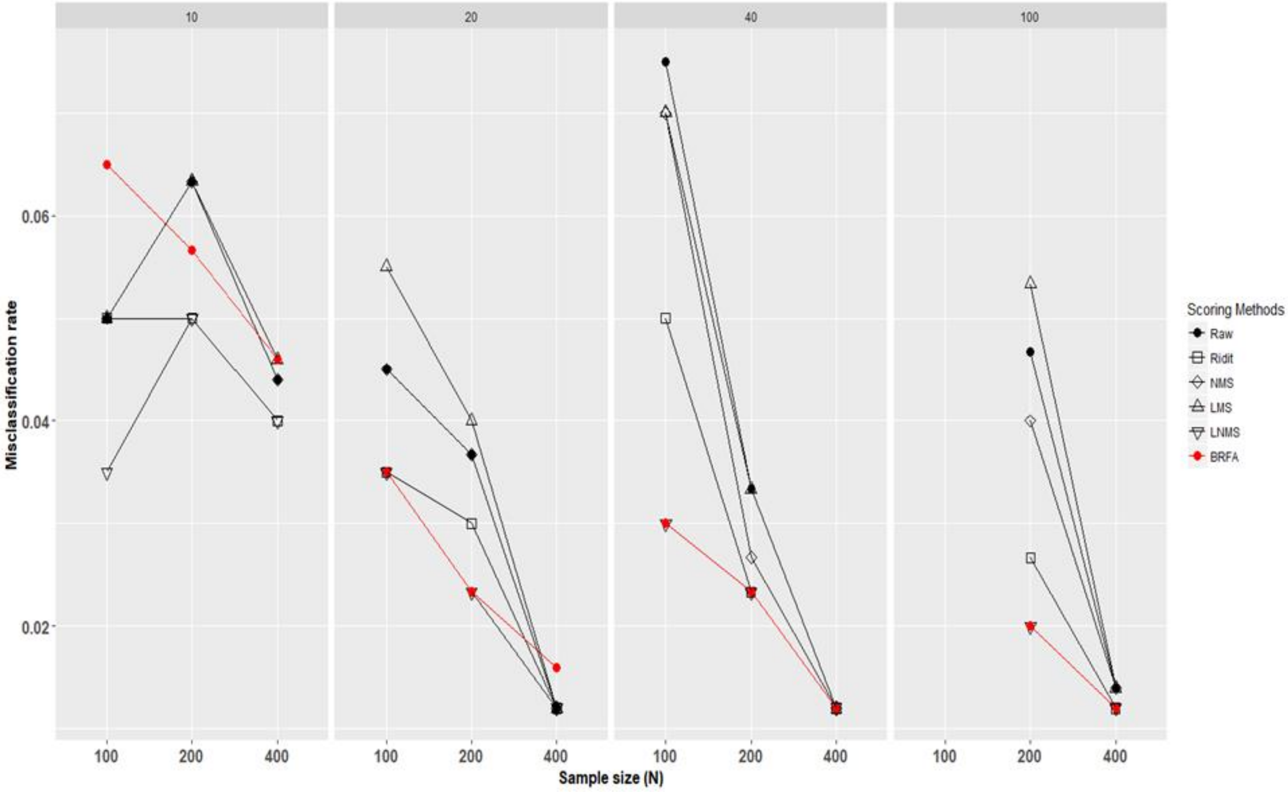

Figure A.14: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

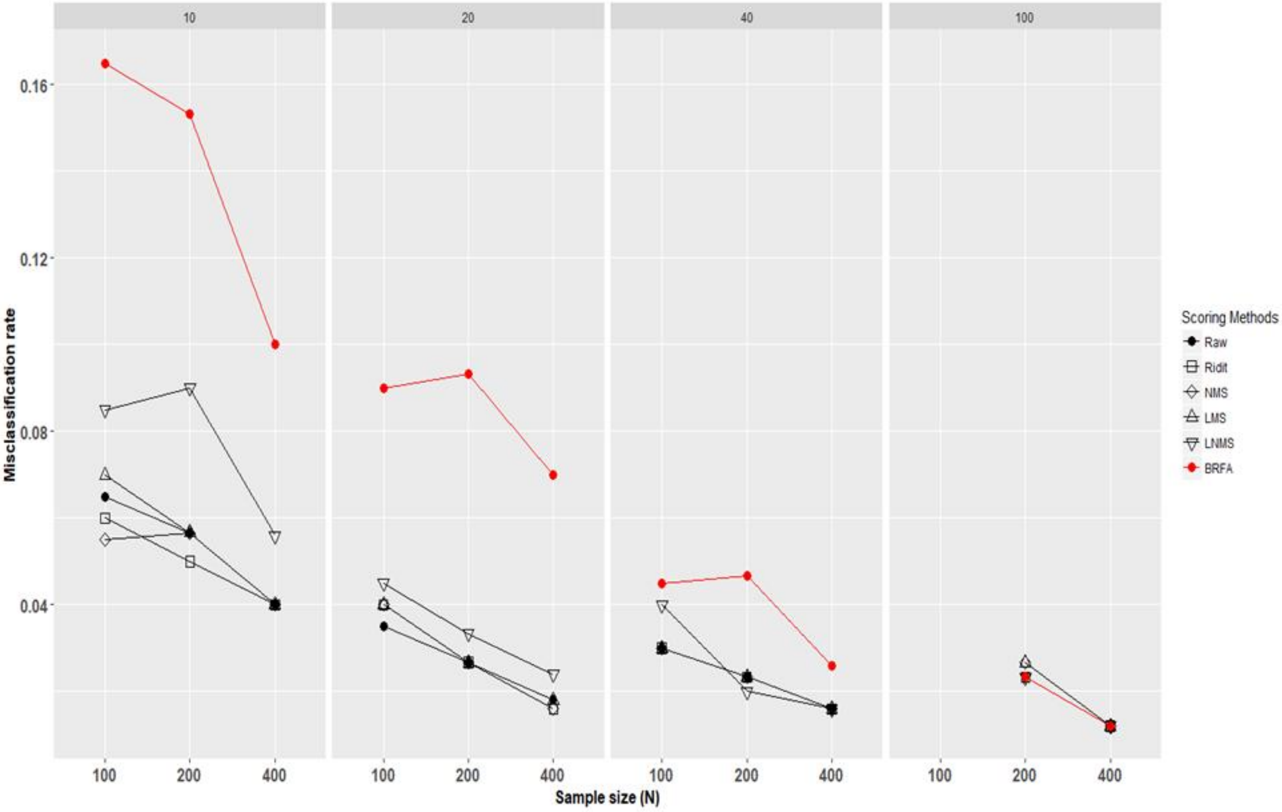

Figure A.15: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

Table A.4: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for LDA in the four considered scenarios. Simulation results refer to the case of ordinal features with three categories ($C = 3$).*

	Scenario 1											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.18 (0.03)	0.24 (0.03)	0.21 (0.02)	0.32 (0.03)	0.25 (0.02)	0.13 (0.01)	0.27 (0.03)	0.15 (0.02)	0.06 (0.01)	-	0.26 (0.03)	0.12 (0.02)
Ridit	0.20 (0.03)	0.24 (0.03)	0.21 (0.02)	0.31 (0.03)	0.24 (0.02)	0.13 (0.01)	0.26 (0.03)	0.16 (0.02)	0.07 (0.01)	-	0.24 (0.03)	0.12 (0.02)
NM	0.19 (0.03)	0.24 (0.04)	0.22 (0.02)	0.32 (0.03)	0.26 (0.03)	0.13 (0.01)	0.28 (0.02)	0.16 (0.02)	0.06 (0.01)	-	0.25 (0.02)	0.12 (0.02)
Blom	0.19 (0.03)	0.24 (0.04)	0.22 (0.02)	0.32 (0.03)	0.26 (0.03)	0.13 (0.01)	0.28 (0.02)	0.16 (0.02)	0.06 (0.01)	-	0.25 (0.02)	0.12 (0.02)
NMS	0.18 (0.03)	0.24 (0.04)	0.22 (0.02)	0.32 (0.03)	0.25 (0.03)	0.13 (0.01)	0.27 (0.02)	0.15 (0.02)	0.07 (0.01)	-	0.25 (0.02)	0.13 (0.02)
LMS	0.19 (0.03)	0.24 (0.04)	0.22 (0.02)	0.34 (0.03)	0.26 (0.03)	0.13 (0.01)	0.28 (0.02)	0.16 (0.02)	0.06 (0.01)	-	0.25 (0.03)	0.12 (0.02)
LNMS	0.30 (0.04)	0.32 (0.02)	0.27 (0.02)	0.34 (0.02)	0.24 (0.02)	0.21 (0.01)	0.29 (0.03)	0.24 (0.03)	0.19 (0.01)	-	0.23 (0.02)	0.19 (0.02)
BRFA	0.23 (0.04)	0.26 (0.02)	0.23 (0.02)	0.23 (0.03)	0.22 (0.02)	0.12 (0.01)	0.09 (0.03)	0.08 (0.02)	0.08 (0.01)	-	0.04 (0.01)	0.02 (0.01)

	Scenario 2											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.04 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.01 (0.01)
Ridit	0.07 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	-	0.03 (0.01)	0.01 (0.01)
NM	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.01 (0.01)
Blom	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.01 (0.01)
NMS	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.01 (0.01)
LMS	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.02 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.01 (0.01)
LNMS	0.16 (0.02)	0.11 (0.02)	0.11 (0.01)	0.14 (0.03)	0.12 (0.01)	0.06 (0.01)	0.20 (0.03)	0.09 (0.01)	0.04 (0.01)	-	0.18 (0.02)	0.07 (0.01)
BRFA	0.14 (0.03)	0.10 (0.02)	0.08 (0.01)	0.10 (0.02)	0.10 (0.02)	0.06 (0.01)	0.08 (0.02)	0.07 (0.01)	0.06 (0.01)	-	0.02 (0.01)	0.02 (0.01)

	Scenario 3											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.01 (0.01)	0.08 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.05 (0.01)	0.05 (0.01)
Ridit	0.05 (0.01)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.01 (0.01)	0.05 (0.02)	0.02 (0.01)	0.01 (0.01)	-	0.03 (0.01)	0.03 (0.01)
NM	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.01 (0.01)	0.07 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.04 (0.01)
Blom	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.01 (0.01)	0.07 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.04 (0.01)
NMS	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.02)	0.04 (0.01)	0.01 (0.01)	0.07 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.04 (0.01)	0.04 (0.01)
LMS	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.06 (0.02)	0.04 (0.01)	0.01 (0.01)	0.07 (0.02)	0.03 (0.01)	0.01 (0.01)	-	0.05 (0.01)	0.05 (0.01)
LNMS	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	-	0.02 (0.01)	0.02 (0.01)
BRFA	0.06 (0.02)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	-	0.02 (0.01)	0.02 (0.01)

	Mixture scenario											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.03)	0.06 (0.02)	0.04 (0.01)	0.04 (0.02)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.02 (0.01)	0.01 (0.01)
Ridit	0.06 (0.02)	0.05 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.02 (0.01)	0.01 (0.01)
NM	0.06 (0.02)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.01 (0.01)
Blom	0.06 (0.02)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.01 (0.01)
NMS	0.06 (0.02)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.01 (0.01)
LMS	0.07 (0.03)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.01 (0.01)
LNMS	0.08 (0.02)	0.09 (0.02)	0.06 (0.01)	0.04 (0.02)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)	-	0.02 (0.01)	0.01 (0.01)
BRFA	0.16 (0.03)	0.15 (0.02)	0.10 (0.01)	0.09 (0.02)	0.09 (0.02)	0.07 (0.01)	0.04 (0.02)	0.05 (0.01)	0.03 (0.01)	-	0.02 (0.01)	0.01 (0.01)

A.2.3 Simulation results for $C = 357$

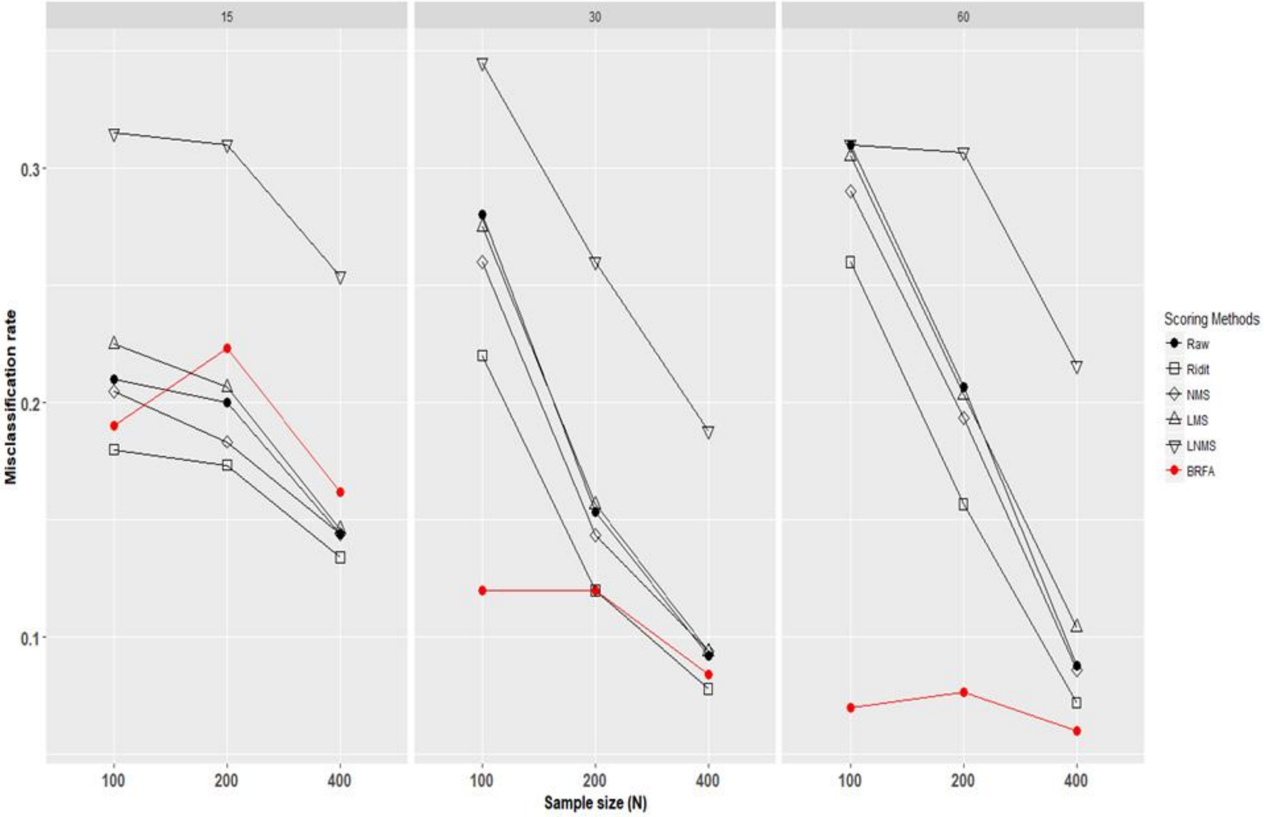

Figure A.16: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

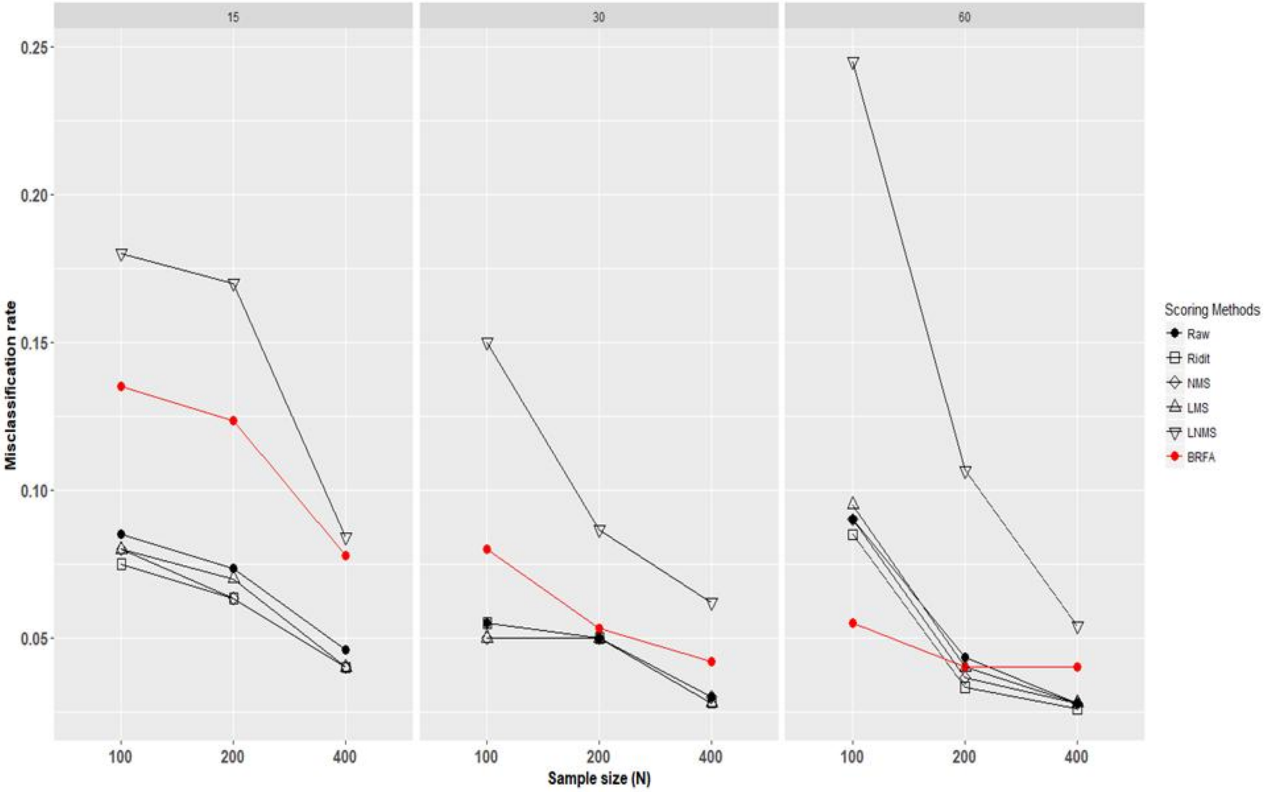

Figure A.17: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$*.*

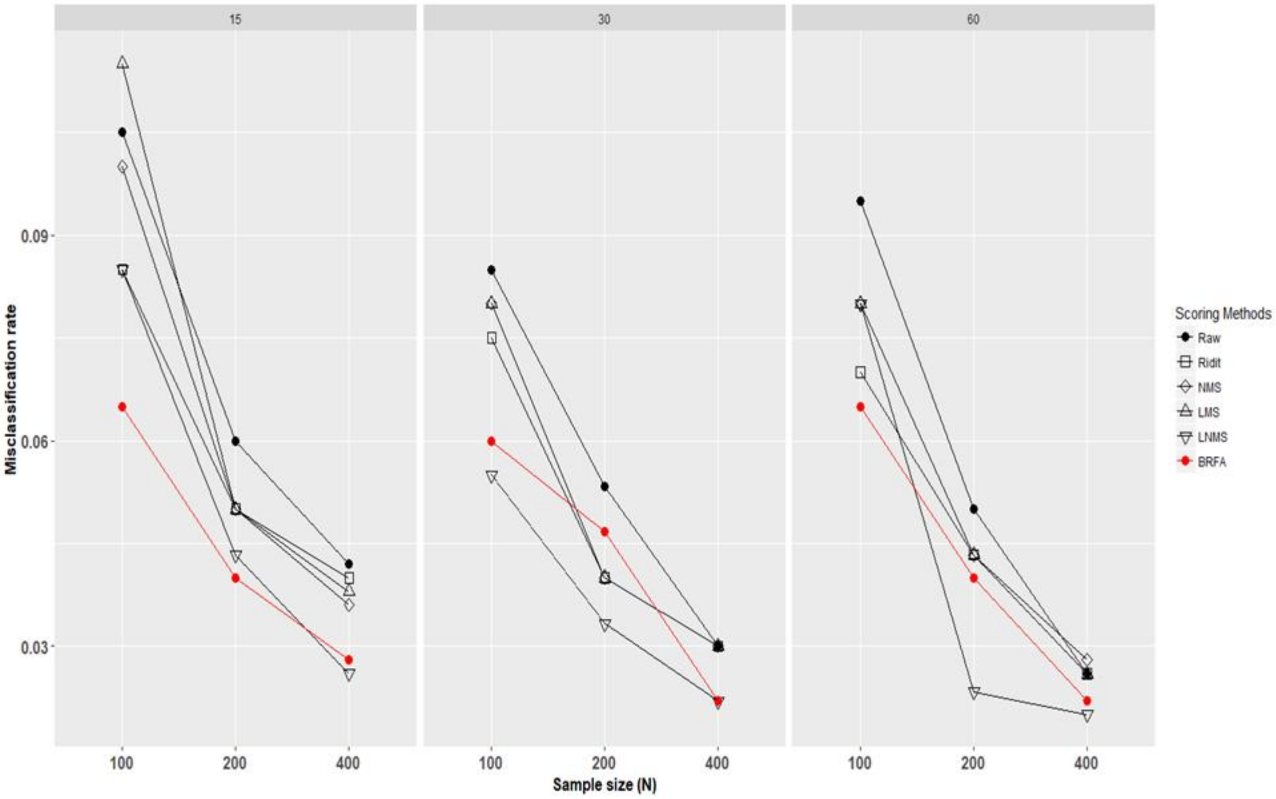

Figure A.18: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

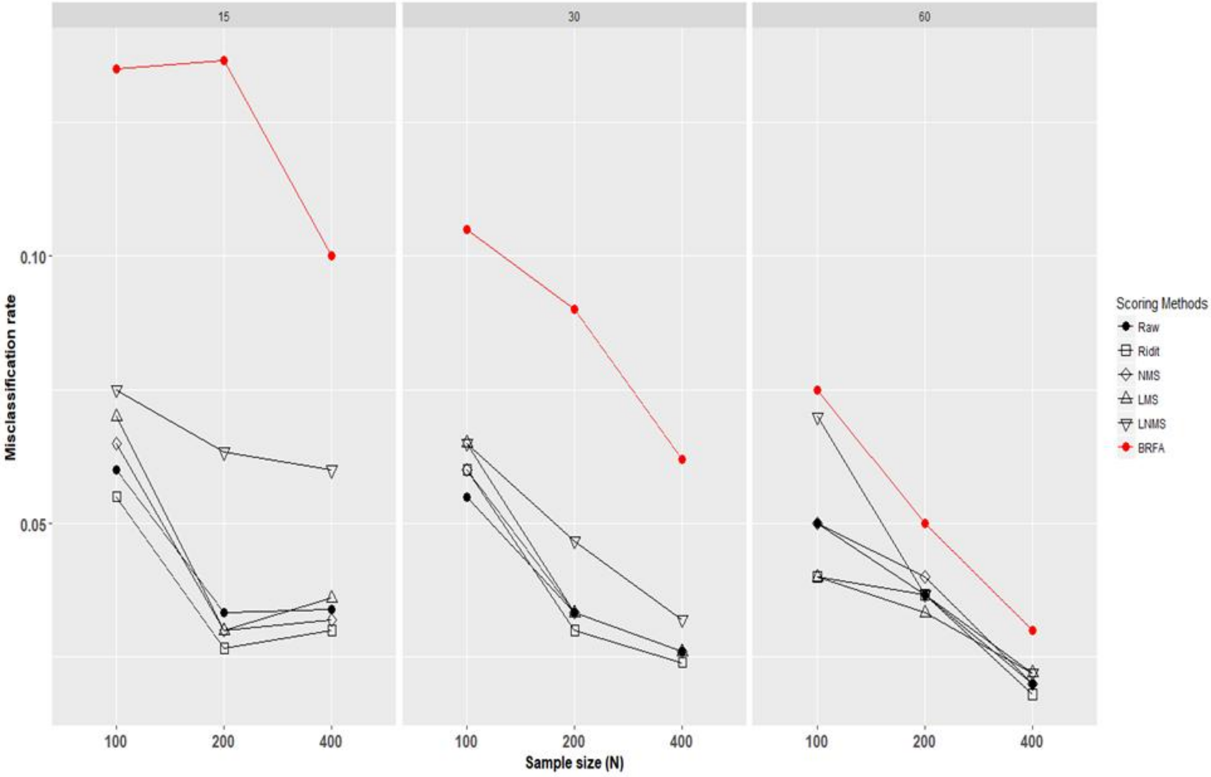

Figure A.19: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

Table A.5: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for LDA in the four considered scenarios. Simulation results refer to the case of ordinal features with different number of categories ($C = 357$).*

	Scenario 1								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.21 (0.03)	0.2 (0.01)	0.14 (0.01)	0.28 (0.03)	0.15 (0.02)	0.09 (0.01)	0.31 (0.02)	0.21 (0.03)	0.09 (0.01)
Ridit	0.18 (0.02)	0.17 (0.01)	0.13 (0.01)	0.22 (0.03)	0.12 (0.02)	0.08 (0.01)	0.26 (0.02)	0.16 (0.02)	0.07 (0.01)
NM	0.20 (0.02)	0.18 (0.01)	0.15 (0.01)	0.26 (0.03)	0.14 (0.02)	0.09 (0.01)	0.29 (0.02)	0.20 (0.02)	0.09 (0.01)
Blom	0.20 (0.02)	0.18 (0.01)	0.15 (0.01)	0.26 (0.03)	0.14 (0.02)	0.09 (0.01)	0.29 (0.02)	0.20 (0.02)	0.09 (0.01)
NMS	0.20 (0.02)	0.18 (0.01)	0.14 (0.01)	0.26 (0.03)	0.14 (0.02)	0.09 (0.01)	0.29 (0.02)	0.19 (0.02)	0.09 (0.01)
LMS	0.22 (0.03)	0.21 (0.01)	0.15 (0.01)	0.28 (0.02)	0.16 (0.02)	0.09 (0.01)	0.3 (0.02)	0.20 (0.02)	0.10 (0.01)
LNMS	0.32 (0.02)	0.31 (0.01)	0.25 (0.01)	0.34 (0.02)	0.26 (0.03)	0.19 (0.02)	0.31 (0.03)	0.31 (0.03)	0.22 (0.02)
BRFA	0.19 (0.03)	0.22 (0.03)	0.16 (0.03)	0.12 (0.02)	0.12 (0.02)	0.08 (0.01)	0.07 (0.02)	0.08 (0.02)	0.06 (0.01)
	Scenario 2								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.03)	0.07 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.09 (0.02)	0.04 (0.01)	0.03 (0.01)
Ridit	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.08 (0.02)	0.03 (0.01)	0.03 (0.01)
NM	0.08 (0.03)	0.07 (0.02)	0.04 (0.01)	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.09 (0.02)	0.04 (0.01)	0.03 (0.01)
Blom	0.08 (0.03)	0.07 (0.02)	0.04 (0.01)	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.09 (0.02)	0.04 (0.01)	0.03 (0.01)
NMS	0.08 (0.03)	0.06 (0.02)	0.04 (0.01)	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.09 (0.02)	0.04 (0.01)	0.03 (0.01)
LMS	0.08 (0.03)	0.07 (0.02)	0.04 (0.01)	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.10 (0.03)	0.04 (0.01)	0.03 (0.01)
LNMS	0.18 (0.03)	0.17 (0.03)	0.08 (0.01)	0.15 (0.02)	0.09 (0.01)	0.06 (0.01)	0.24 (0.03)	0.11 (0.01)	0.05 (0.01)
BRFA	0.14 (0.02)	0.12 (0.02)	0.08 (0.02)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)
	Scenario 3								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.10 (0.02)	0.06 (0.01)	0.04 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.10 (0.03)	0.05 (0.01)	0.03 (0.01)
Ridit	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)
NM	0.10 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
Blom	0.10 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
NMS	0.10 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)
LMS	0.12 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)
LNMS	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.02 (0.01)	0.08 (0.02)	0.02 (0.01)	0.02 (0.01)
BRFA	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
	Mixture scenario								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
LMS	0.07 (0.02)	0.03 (0.01)	0.04 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
LNMS	0.08 (0.03)	0.06 (0.01)	0.06 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
BRFA	0.14 (0.02)	0.14 (0.02)	0.10 (0.02)	0.10 (0.03)	0.09 (0.01)	0.06 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)

A.3 QDA

The QDA, along with the LDA, is the classifier which returns the worst performance on the simulated data in each scenario, whatever the value of C is. In figures from A.20 to A.26 it is possible to notice that, when $p = 100$, the classifier applied on BRFA performs much better if compared to any other scoring method and the gain in term of correctly classified observations is really remarkable. When $p = 10$ or $p = 20$ the misclassification rates associated with the BRFA method do not significantly differ from the other scoring methods in the majority of the situations, though when $p = 20$ the BRFA is often the optimal one. When $p = 40$ the BRFA returns better results only if the number of instances N is 100 or 200. When $N = 400$ the gap between the misclassification rates associated with the scoring methods is significantly smaller.

In the case of simulated datasets composed by features with different number of categories (i.e. $C = 357$) there is a clear advantage in using the BRFA in any scenario when the number of features is high (i.e. $p = 60$) and the number of instances is 100 or 200.

In scenario 1 the BRFA mean misclassification rate is lower than any other when $p = 30$ or $p = 60$ while, when $p = 15$, for every values of N it is the best or second best. In scenario 2 and 3 when $p = 15$ the BRFA mean misclassification rates do not significantly differ from the other scoring methods. The same situation occurs when $p = 30$, with the only exception of the case $N = 100$, where the BRFA mean misclassification rate is the best one.

A.3.1 Simulation results for $C = 5$

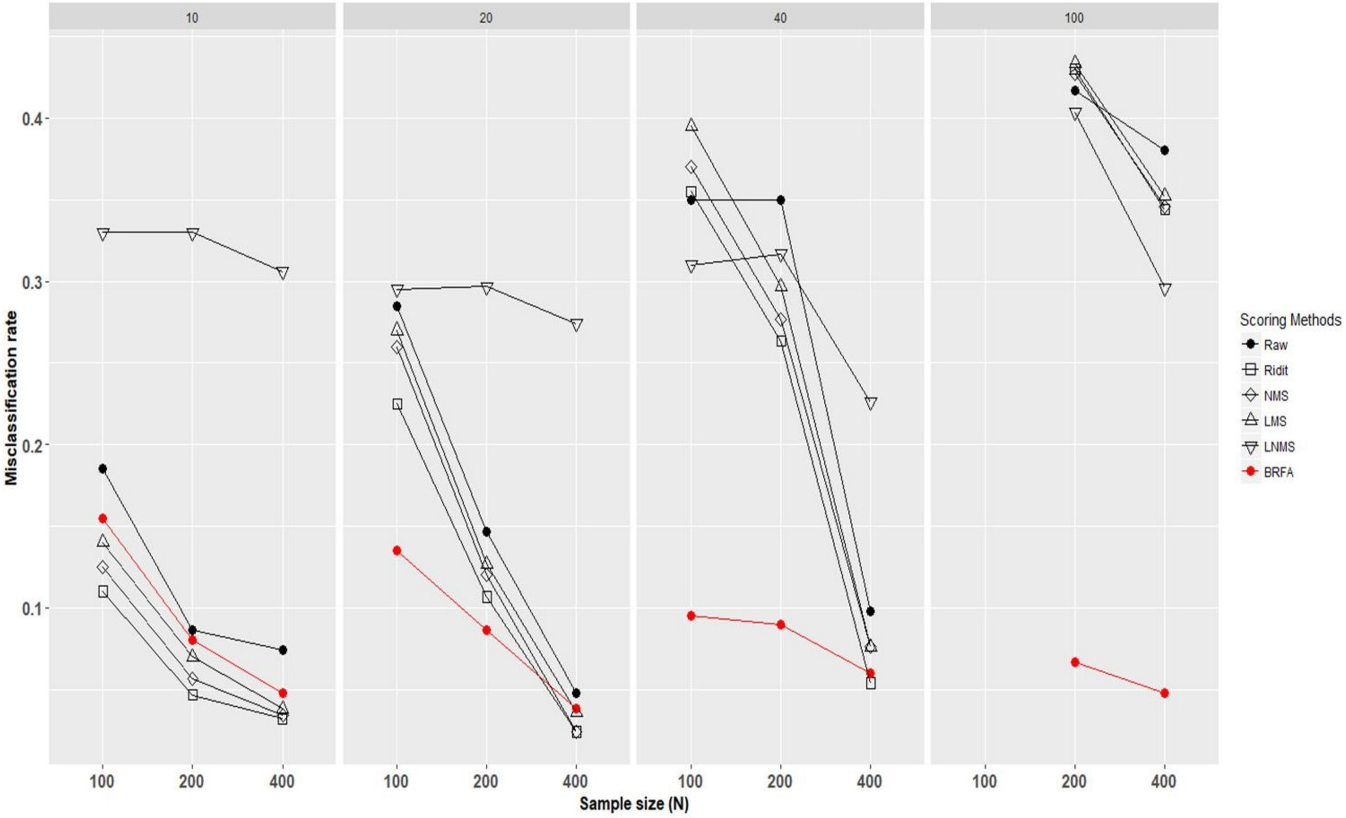

Figure A.20: *Results for Scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

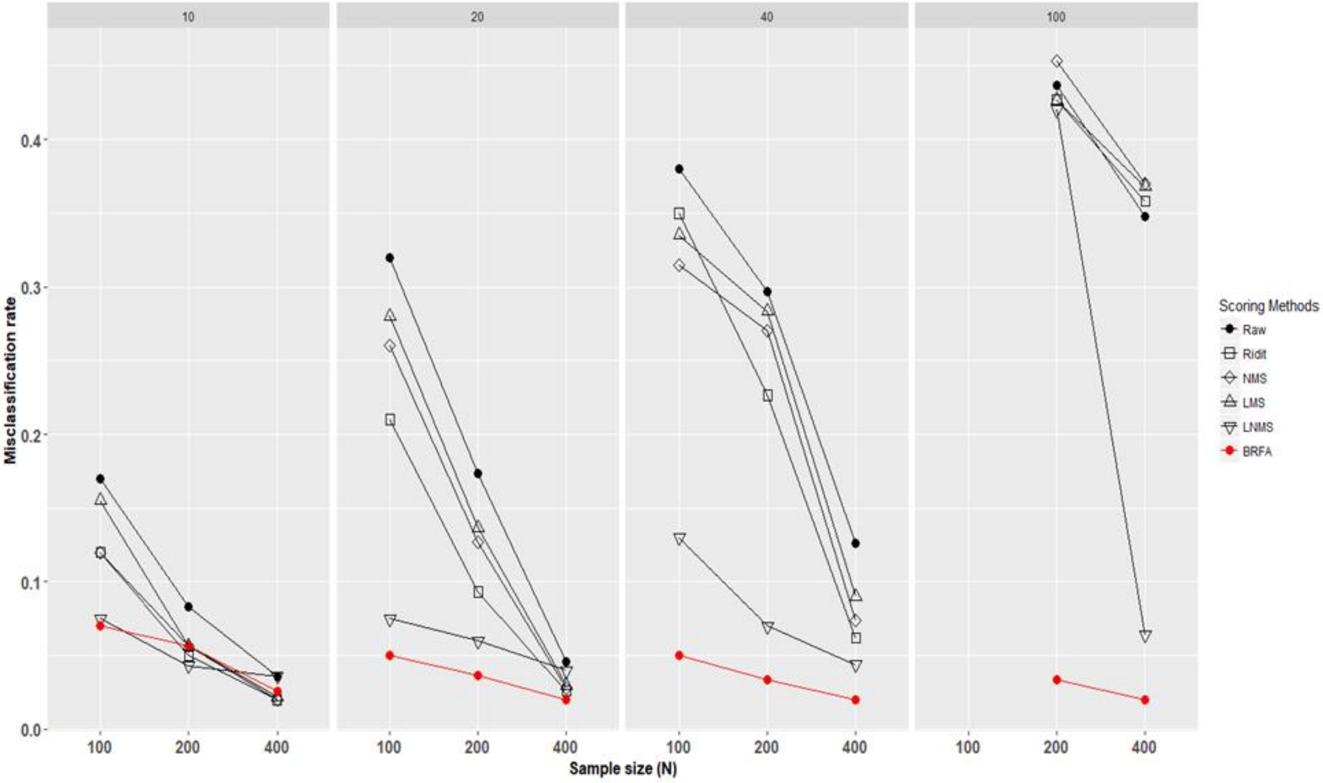

Figure A.21: *Results for Scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

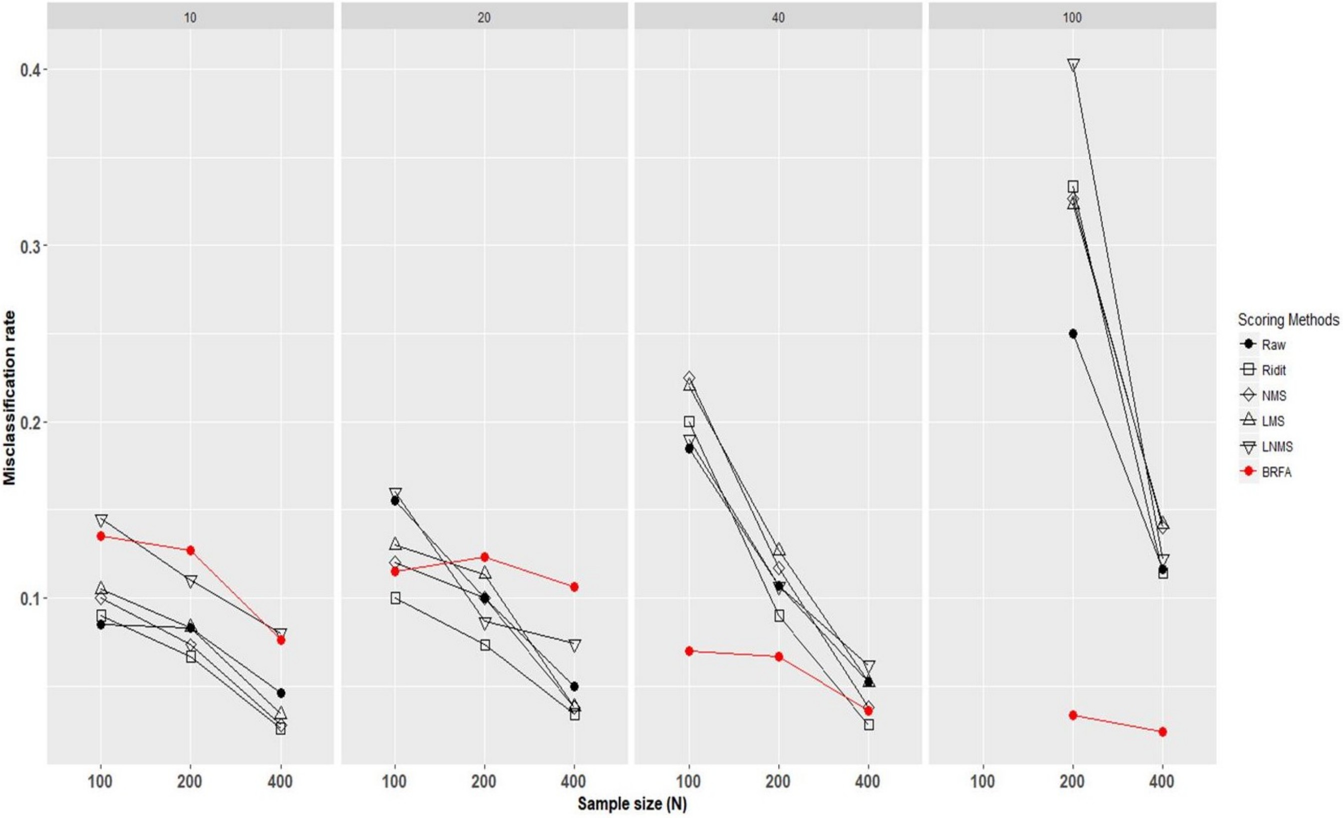

Figure A.22: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

Table A.6: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for QDA in the four considered scenarios. Simulation results refer to the case of ordinal features with five categories ($C = 5$).*

	Scenario 1											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.40 (0.02)	0.32 (0.02)	0.24 (0.02)	0.40 (0.02)	0.36 (0.03)	0.27 (0.03)	0.42 (0.02)	0.40 (0.02)	0.32 (0.02)	-	0.39 (0.02)	0.45 (0.01)
Ridit	0.31 (0.03)	0.27 (0.01)	0.21 (0.01)	0.38 (0.02)	0.31 (0.03)	0.24 (0.02)	0.4 (0.01)	0.37 (0.02)	0.29 (0.02)	-	0.43 (0.02)	0.45 (0.01)
NM	0.36 (0.02)	0.32 (0.03)	0.24 (0.01)	0.40 (0.02)	0.35 (0.03)	0.28 (0.03)	0.42 (0.02)	0.39 (0.02)	0.31 (0.02)	-	0.40 (0.02)	0.44 (0.01)
Blom	0.36 (0.02)	0.32 (0.03)	0.24 (0.01)	0.40 (0.02)	0.35 (0.03)	0.28 (0.03)	0.43 (0.01)	0.39 (0.02)	0.31 (0.02)	-	0.40 (0.02)	0.44 (0.01)
NMS	0.36 (0.02)	0.31 (0.03)	0.24 (0.02)	0.40 (0.02)	0.35 (0.03)	0.27 (0.03)	0.42 (0.02)	0.39 (0.01)	0.32 (0.02)	-	0.41 (0.02)	0.44 (0.01)
LMS	0.38 (0.02)	0.33 (0.03)	0.25 (0.02)	0.42 (0.02)	0.35 (0.03)	0.30 (0.03)	0.42 (0.01)	0.40 (0.03)	0.32 (0.02)	-	0.40 (0.02)	0.43 (0.01)
LNMS	0.40 (0.02)	0.36 (0.02)	0.40 (0.02)	0.40 (0.02)	0.36 (0.03)	0.39 (0.02)	0.34 (0.03)	0.34 (0.03)	0.35 (0.02)	-	0.41 (0.02)	0.33 (0.01)
BRFA	0.22 (0.03)	0.24 (0.03)	0.23 (0.03)	0.20 (0.03)	0.22 (0.03)	0.15 (0.01)	0.06 (0.01)	0.08 (0.02)	0.07 (0.01)	-	0.05 (0.01)	0.04 (0.01)
	Scenario 2											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.18 (0.02)	0.09 (0.02)	0.07 (0.01)	0.28 (0.02)	0.15 (0.02)	0.05 (0.01)	0.35 (0.03)	0.35 (0.02)	0.10 (0.01)	-	0.42 (0.02)	0.38 (0.02)
Ridit	0.11 (0.01)	0.05 (0.01)	0.03 (0.01)	0.22 (0.02)	0.11 (0.01)	0.02 (0.01)	0.36 (0.02)	0.26 (0.03)	0.05 (0.01)	-	0.43 (0.02)	0.34 (0.02)
NM	0.12 (0.02)	0.06 (0.01)	0.03 (0.01)	0.25 (0.02)	0.12 (0.01)	0.03 (0.01)	0.39 (0.02)	0.28 (0.02)	0.08 (0.01)	-	0.44 (0.01)	0.35 (0.02)
Blom	0.12 (0.02)	0.06 (0.01)	0.03 (0.01)	0.25 (0.02)	0.12 (0.01)	0.03 (0.01)	0.39 (0.02)	0.28 (0.02)	0.08 (0.01)	-	0.44 (0.01)	0.35 (0.02)
NMS	0.12 (0.03)	0.06 (0.01)	0.03 (0.01)	0.26 (0.02)	0.12 (0.01)	0.02 (0.01)	0.37 (0.02)	0.28 (0.02)	0.08 (0.01)	-	0.43 (0.02)	0.35 (0.02)
LMS	0.14 (0.02)	0.07 (0.01)	0.04 (0.01)	0.27 (0.02)	0.13 (0.01)	0.04 (0.01)	0.40 (0.02)	0.30 (0.02)	0.08 (0.01)	-	0.43 (0.01)	0.35 (0.01)
LNMS	0.33 (0.04)	0.33 (0.02)	0.31 (0.01)	0.30 (0.02)	0.30 (0.03)	0.27 (0.03)	0.31 (0.02)	0.32 (0.02)	0.23 (0.01)	-	0.40 (0.02)	0.30 (0.02)
BRFA	0.16 (0.03)	0.08 (0.01)	0.05 (0.01)	0.14 (0.03)	0.09 (0.02)	0.04 (0.01)	0.10 (0.02)	0.09 (0.01)	0.06 (0.01)	-	0.07 (0.01)	0.05 (0.01)
	Scenario 3											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.17 (0.03)	0.08 (0.01)	0.04 (0.01)	0.32 (0.04)	0.17 (0.01)	0.05 (0.01)	0.38 (0.03)	0.30 (0.02)	0.13 (0.01)	-	0.44 (0.02)	0.35 (0.02)
Ridit	0.12 (0.02)	0.05 (0.01)	0.02 (0.01)	0.21 (0.04)	0.09 (0.02)	0.03 (0.01)	0.35 (0.03)	0.23 (0.02)	0.06 (0.01)	-	0.43 (0.02)	0.36 (0.02)
NM	0.14 (0.03)	0.06 (0.01)	0.02 (0.01)	0.27 (0.04)	0.14 (0.02)	0.03 (0.01)	0.35 (0.03)	0.29 (0.02)	0.08 (0.01)	-	0.44 (0.01)	0.37 (0.02)
Blom	0.14 (0.03)	0.06 (0.01)	0.02 (0.01)	0.27 (0.04)	0.14 (0.02)	0.03 (0.01)	0.35 (0.03)	0.29 (0.02)	0.08 (0.01)	-	0.44 (0.01)	0.37 (0.02)
NMS	0.12 (0.03)	0.06 (0.01)	0.02 (0.01)	0.26 (0.03)	0.13 (0.02)	0.03 (0.01)	0.32 (0.03)	0.27 (0.02)	0.07 (0.01)	-	0.45 (0.01)	0.37 (0.01)
LMS	0.16 (0.03)	0.06 (0.01)	0.02 (0.01)	0.28 (0.03)	0.14 (0.02)	0.03 (0.01)	0.34 (0.02)	0.28 (0.02)	0.09 (0.01)	-	0.43 (0.01)	0.37 (0.01)
LNMS	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.13 (0.03)	0.07 (0.01)	0.04 (0.01)	-	0.42 (0.02)	0.06 (0.02)
BRFA	0.07 (0.02)	0.06 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.02 (0.01)
	Mixture scenario											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.08 (0.02)	0.05 (0.01)	0.16 (0.02)	0.10 (0.01)	0.05 (0.01)	0.18 (0.03)	0.11 (0.02)	0.05 (0.01)	-	0.25 (0.03)	0.12 (0.01)
Ridit	0.09 (0.02)	0.07 (0.02)	0.03 (0.01)	0.10 (0.02)	0.07 (0.01)	0.03 (0.01)	0.20 (0.03)	0.09 (0.01)	0.03 (0.01)	-	0.33 (0.01)	0.11 (0.01)
NM	0.10 (0.02)	0.08 (0.02)	0.03 (0.01)	0.14 (0.02)	0.1 (0.01)	0.04 (0.01)	0.23 (0.04)	0.11 (0.01)	0.04 (0.01)	-	0.31 (0.01)	0.14 (0.01)
Blom	0.10 (0.02)	0.08 (0.02)	0.03 (0.01)	0.14 (0.02)	0.11 (0.01)	0.04 (0.01)	0.23 (0.04)	0.11 (0.01)	0.04 (0.01)	-	0.31 (0.01)	0.14 (0.01)
NMS	0.10 (0.02)	0.07 (0.02)	0.03 (0.01)	0.12 (0.02)	0.10 (0.01)	0.04 (0.01)	0.22 (0.03)	0.12 (0.01)	0.04 (0.01)	-	0.33 (0.02)	0.14 (0.01)
LMS	0.10 (0.02)	0.08 (0.02)	0.03 (0.01)	0.13 (0.02)	0.11 (0.01)	0.04 (0.01)	0.22 (0.03)	0.13 (0.01)	0.05 (0.01)	-	0.32 (0.02)	0.14 (0.01)
LNMS	0.14 (0.03)	0.11 (0.02)	0.08 (0.01)	0.16 (0.02)	0.09 (0.01)	0.07 (0.01)	0.19 (0.02)	0.11 (0.02)	0.06 (0.01)	-	0.40 (0.03)	0.12 (0.02)
BRFA	0.14 (0.02)	0.13 (0.02)	0.08 (0.01)	0.12 (0.03)	0.12 (0.02)	0.11 (0.01)	0.07 (0.02)	0.07 (0.01)	0.04 (0.01)	-	0.03 (0.01)	0.02 (0.01)

A.3.2 Simulation results for $C = 3$

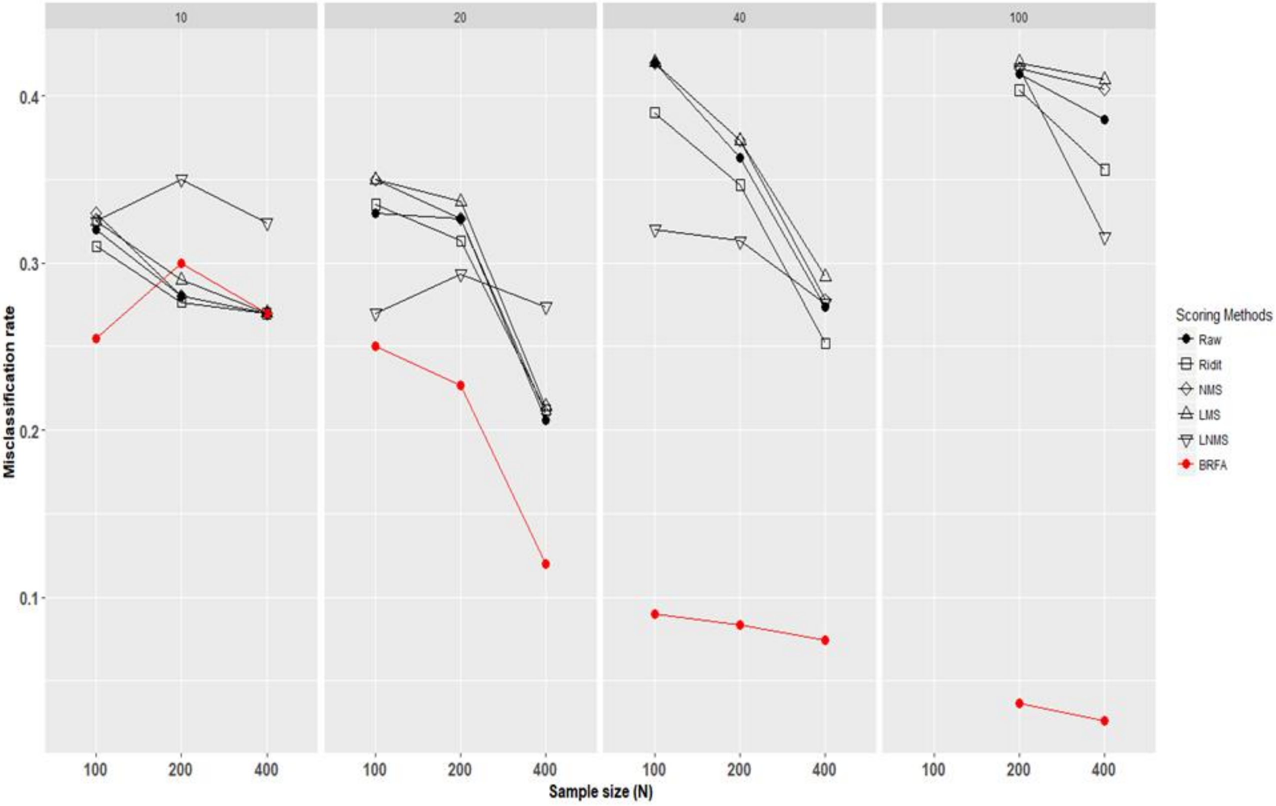

Figure A.23: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

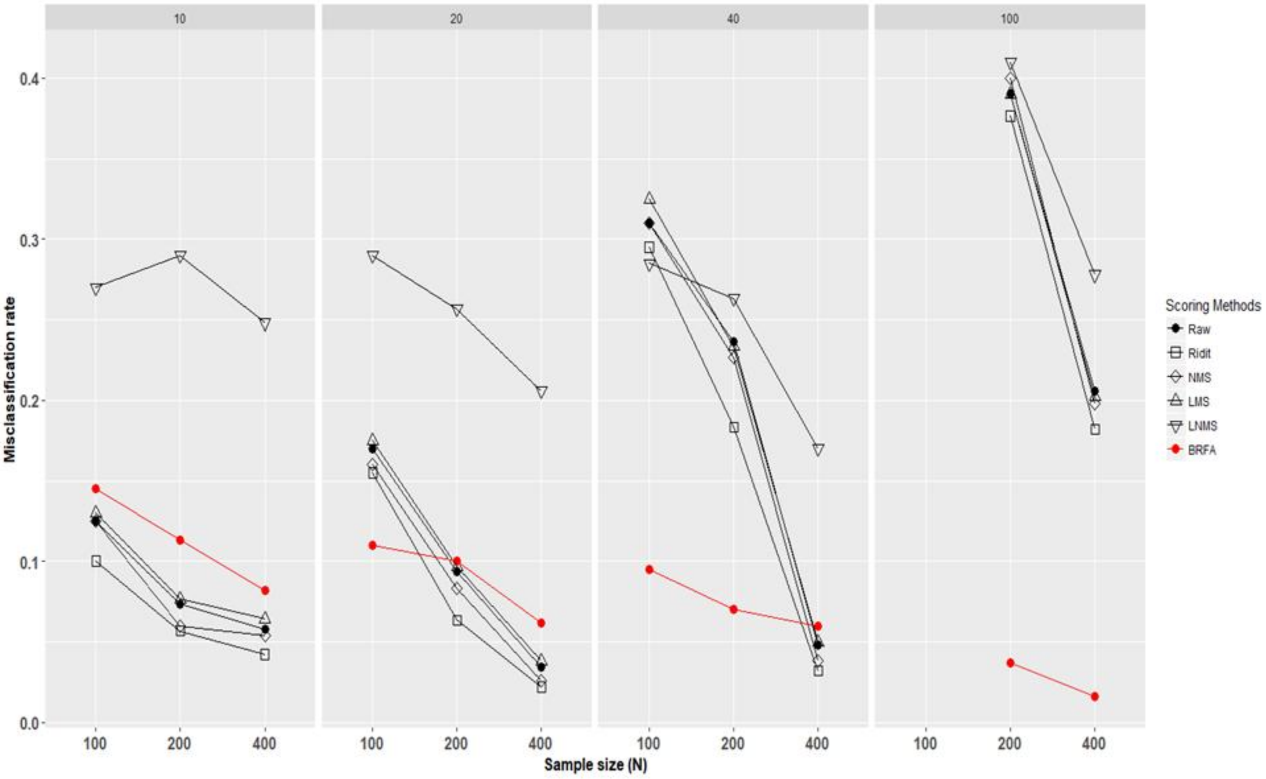

Figure A.24: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

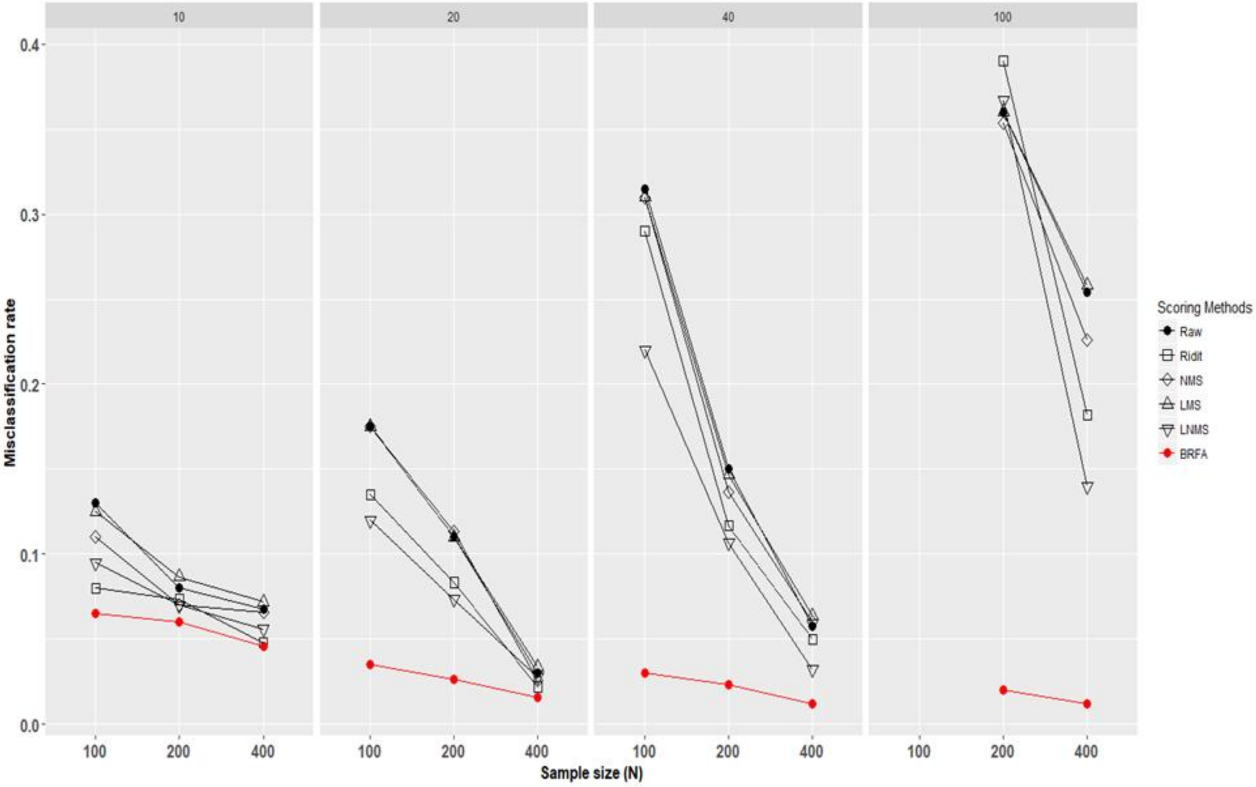

Figure A.25: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

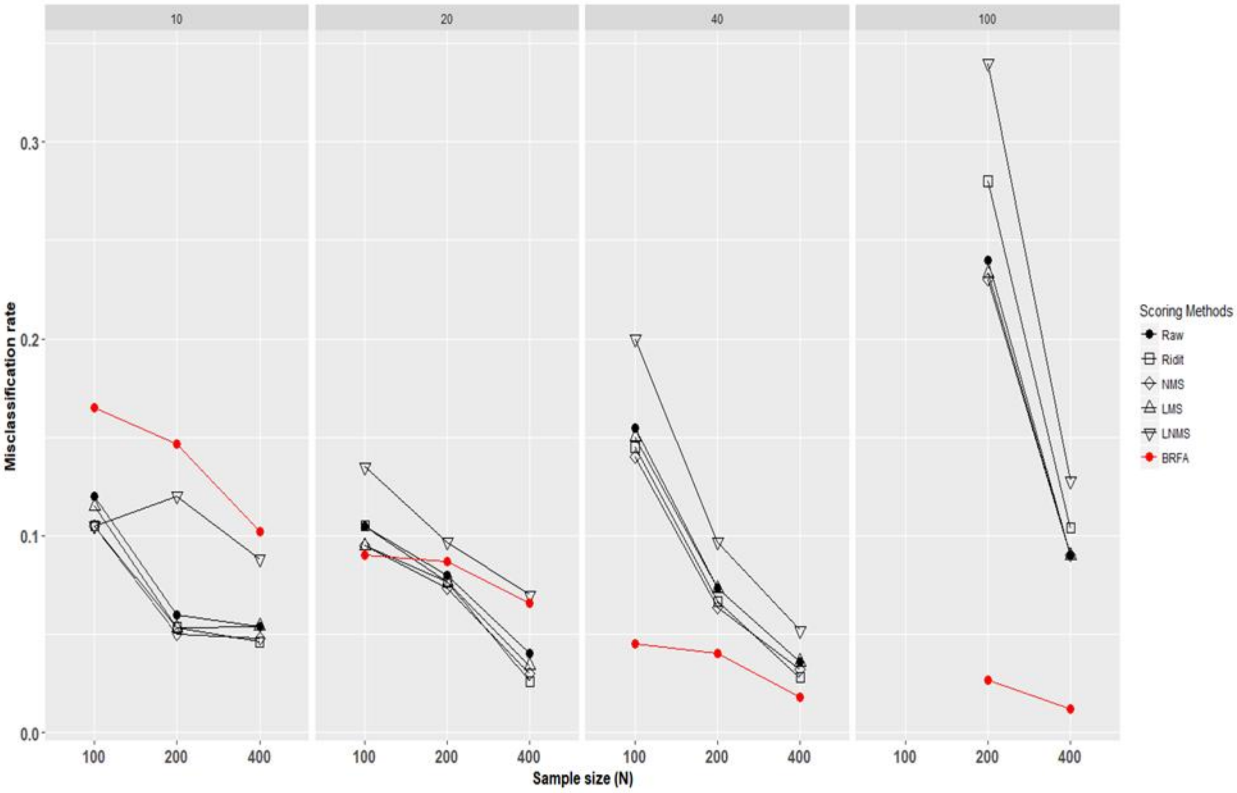

Figure A.26: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

Table A.7: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for QDA in the four considered scenarios. Simulation results refer to the case of ordinal features with three categories (C = 3).*

						Scenario 1						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.32 (0.03)	0.28 (0.03)	0.27 (0.02)	0.33 (0.03)	0.33 (0.02)	0.21 (0.02)	0.42 (0.01)	0.36 (0.02)	0.27 (0.02)	-	0.41 (0.02)	0.39 (0.02)
Ridit	0.31 (0.03)	0.28 (0.03)	0.27 (0.02)	0.34 (0.04)	0.31 (0.02)	0.21 (0.01)	0.39 (0.02)	0.35 (0.02)	0.25 (0.03)	-	0.40 (0.02)	0.36 (0.02)
NM	0.33 (0.03)	0.28 (0.02)	0.27 (0.02)	0.34 (0.03)	0.34 (0.02)	0.22 (0.01)	0.42 (0.01)	0.37 (0.01)	0.29 (0.03)	-	0.42 (0.02)	0.41 (0.02)
Blom	0.33 (0.03)	0.28 (0.02)	0.27 (0.02)	0.34 (0.03)	0.34 (0.02)	0.22 (0.01)	0.42 (0.01)	0.37 (0.01)	0.29 (0.03)	-	0.42 (0.02)	0.41 (0.02)
NMS	0.33 (0.03)	0.28 (0.03)	0.27 (0.02)	0.35 (0.03)	0.33 (0.02)	0.21 (0.02)	0.42 (0.01)	0.37 (0.02)	0.28 (0.02)	-	0.42 (0.02)	0.40 (0.02)
LMS	0.32 (0.03)	0.29 (0.02)	0.27 (0.02)	0.35 (0.03)	0.34 (0.02)	0.21 (0.02)	0.42 (0.01)	0.37 (0.01)	0.29 (0.03)	-	0.42 (0.02)	0.41 (0.02)
LNMS	0.32 (0.04)	0.35 (0.02)	0.32 (0.01)	0.27 (0.01)	0.29 (0.03)	0.27 (0.02)	0.32 (0.03)	0.31 (0.02)	0.28 (0.02)	-	0.42 (0.02)	0.32 (0.02)
BRFA	0.26 (0.04)	0.30 (0.02)	0.27 (0.02)	0.25 (0.03)	0.23 (0.02)	0.12 (0.01)	0.09 (0.03)	0.08 (0.02)	0.07 (0.01)	-	0.04 (0.01)	0.03 (0.01)
						Scenario 2						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.12 (0.03)	0.07 (0.01)	0.06 (0.01)	0.17 (0.02)	0.09 (0.02)	0.03 (0.01)	0.31 (0.03)	0.24 (0.02)	0.05 (0.01)	-	0.39 (0.02)	0.21 (0.02)
Ridit	0.10 (0.03)	0.06 (0.01)	0.04 (0.01)	0.16 (0.04)	0.06 (0.01)	0.02 (0.01)	0.30 (0.03)	0.18 (0.02)	0.03 (0.01)	-	0.38 (0.02)	0.18 (0.02)
NM	0.12 (0.03)	0.07 (0.01)	0.06 (0.01)	0.17 (0.02)	0.08 (0.01)	0.03 (0.01)	0.32 (0.03)	0.23 (0.02)	0.05 (0.01)	-	0.39 (0.02)	0.21 (0.02)
Blom	0.12 (0.03)	0.07 (0.01)	0.06 (0.01)	0.17 (0.02)	0.08 (0.01)	0.03 (0.01)	0.32 (0.03)	0.23 (0.02)	0.04 (0.01)	-	0.39 (0.02)	0.21 (0.02)
NMS	0.12 (0.03)	0.06 (0.01)	0.05 (0.01)	0.16 (0.02)	0.08 (0.01)	0.03 (0.01)	0.31 (0.04)	0.23 (0.01)	0.04 (0.01)	-	0.40 (0.02)	0.20 (0.02)
LMS	0.13 (0.03)	0.08 (0.01)	0.06 (0.01)	0.18 (0.02)	0.10 (0.02)	0.04 (0.01)	0.32 (0.03)	0.23 (0.02)	0.05 (0.01)	-	0.39 (0.02)	0.20 (0.02)
LNMS	0.27 (0.03)	0.29 (0.04)	0.25 (0.02)	0.29 (0.04)	0.26 (0.02)	0.21 (0.02)	0.28 (0.03)	0.26 (0.02)	0.17 (0.01)	-	0.41 (0.02)	0.28 (0.02)
BRFA	0.14 (0.03)	0.11 (0.02)	0.08 (0.01)	0.11 (0.02)	0.10 (0.02)	0.06 (0.01)	0.10 (0.02)	0.07 (0.01)	0.06 (0.01)	-	0.04 (0.01)	0.02 (0.01)
						Scenario 3						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.13 (0.03)	0.08 (0.01)	0.07 (0.01)	0.18 (0.03)	0.11 (0.02)	0.03 (0.01)	0.32 (0.04)	0.15 (0.03)	0.06 (0.01)	-	0.36 (0.02)	0.36 (0.02)
Ridit	0.08 (0.02)	0.07 (0.01)	0.05 (0.01)	0.14 (0.03)	0.08 (0.01)	0.02 (0.01)	0.29 (0.03)	0.12 (0.03)	0.05 (0.01)	-	0.39 (0.02)	0.39 (0.02)
NM	0.12 (0.03)	0.07 (0.01)	0.07 (0.01)	0.18 (0.03)	0.11 (0.02)	0.03 (0.01)	0.31 (0.04)	0.15 (0.03)	0.06 (0.01)	-	0.35 (0.02)	0.35 (0.02)
Blom	0.11 (0.03)	0.07 (0.01)	0.07 (0.01)	0.18 (0.03)	0.11 (0.02)	0.03 (0.01)	0.31 (0.04)	0.15 (0.03)	0.06 (0.01)	-	0.35 (0.02)	0.35 (0.02)
NMS	0.11 (0.03)	0.07 (0.01)	0.07 (0.01)	0.18 (0.03)	0.11 (0.02)	0.03 (0.01)	0.31 (0.03)	0.14 (0.02)	0.06 (0.01)	-	0.35 (0.01)	0.35 (0.01)
LMS	0.12 (0.03)	0.09 (0.01)	0.07 (0.01)	0.18 (0.03)	0.11 (0.02)	0.03 (0.01)	0.31 (0.03)	0.15 (0.03)	0.06 (0.01)	-	0.36 (0.02)	0.36 (0.02)
LNMS	0.10 (0.02)	0.07 (0.01)	0.06 (0.01)	0.12 (0.03)	0.07 (0.01)	0.03 (0.01)	0.22 (0.04)	0.11 (0.02)	0.03 (0.01)	-	0.37 (0.01)	0.37 (0.01)
BRFA	0.06 (0.02)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	-	0.02 (0.01)	0.02 (0.01)
						Mixture scenario						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.12 (0.03)	0.06 (0.01)	0.05 (0.01)	0.10 (0.03)	0.08 (0.02)	0.04 (0.01)	0.16 (0.02)	0.07 (0.01)	0.04 (0.01)	-	0.24 (0.03)	0.09 (0.01)
Ridit	0.1 (0.03)	0.05 (0.02)	0.05 (0.01)	0.10 (0.03)	0.08 (0.01)	0.03 (0.01)	0.14 (0.02)	0.07 (0.01)	0.03 (0.01)	-	0.28 (0.03)	0.10 (0.02)
NM	0.11 (0.03)	0.05 (0.01)	0.05 (0.01)	0.10 (0.03)	0.07 (0.02)	0.03 (0.01)	0.14 (0.02)	0.07 (0.01)	0.03 (0.01)	-	0.22 (0.03)	0.09 (0.01)
Blom	0.11 (0.03)	0.05 (0.01)	0.05 (0.01)	0.10 (0.03)	0.07 (0.02)	0.03 (0.01)	0.14 (0.02)	0.07 (0.01)	0.03 (0.01)	-	0.22 (0.03)	0.09 (0.01)
NMS	0.1 (0.03)	0.05 (0.01)	0.05 (0.01)	0.10 (0.03)	0.07 (0.02)	0.03 (0.01)	0.14 (0.02)	0.06 (0.01)	0.03 (0.01)	-	0.23 (0.03)	0.09 (0.01)
LMS	0.12 (0.03)	0.05 (0.01)	0.05 (0.01)	0.10 (0.03)	0.08 (0.02)	0.03 (0.01)	0.15 (0.02)	0.07 (0.01)	0.04 (0.01)	-	0.23 (0.03)	0.09 (0.01)
LNMS	0.10 (0.02)	0.12 (0.02)	0.09 (0.01)	0.14 (0.02)	0.10 (0.02)	0.07 (0.01)	0.20 (0.03)	0.10 (0.01)	0.05 (0.01)	-	0.34 (0.04)	0.13 (0.02)
BRFA	0.16 (0.03)	0.15 (0.02)	0.10 (0.01)	0.09 (0.02)	0.09 (0.02)	0.07 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	-	0.03 (0.01)	0.01 (0.01)

A.3.3 Simulation results for $C = 357$

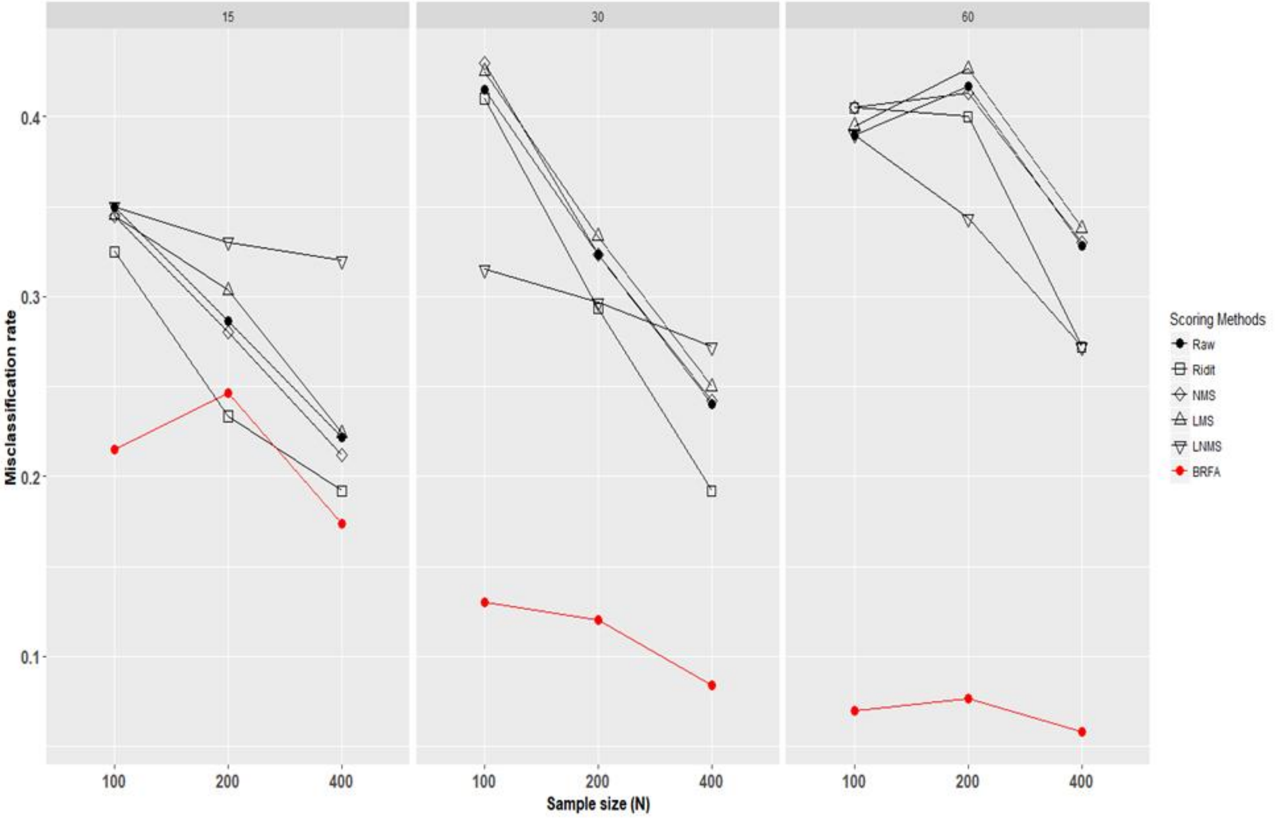

Figure A.27: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$*.*

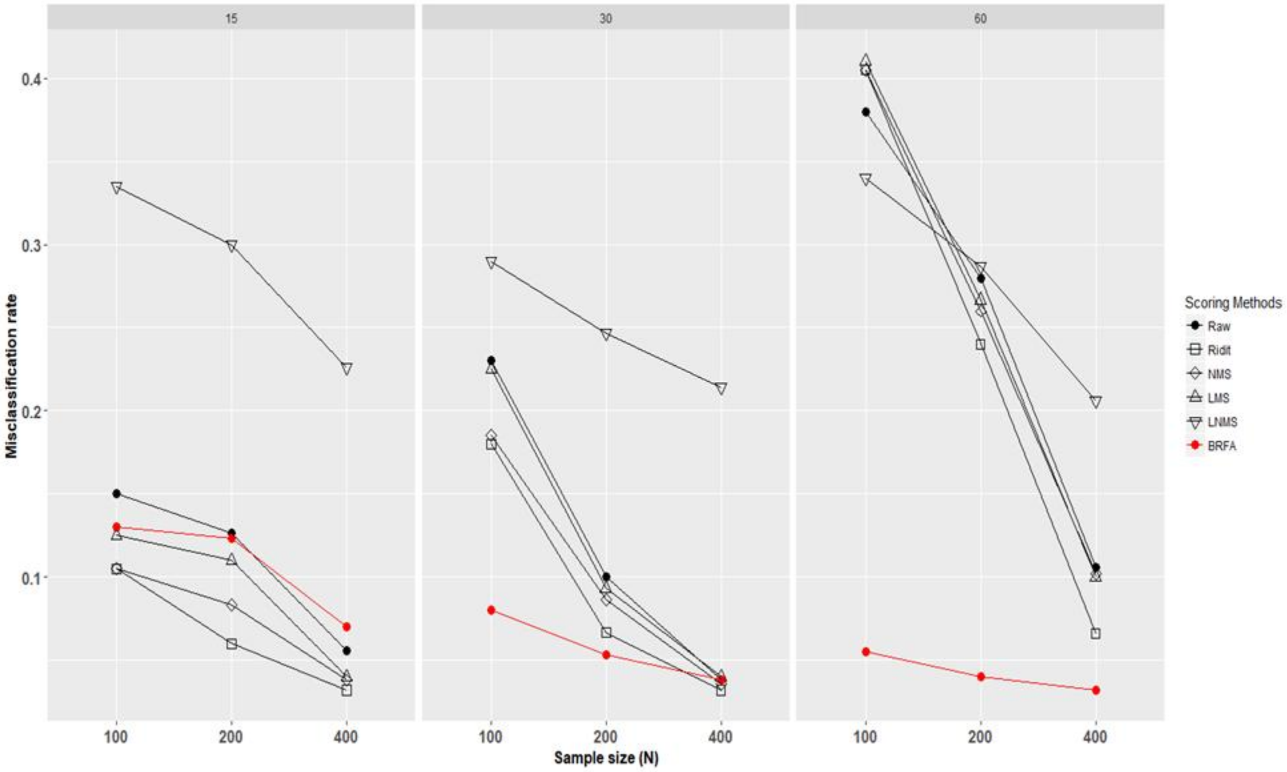

Figure A.28: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

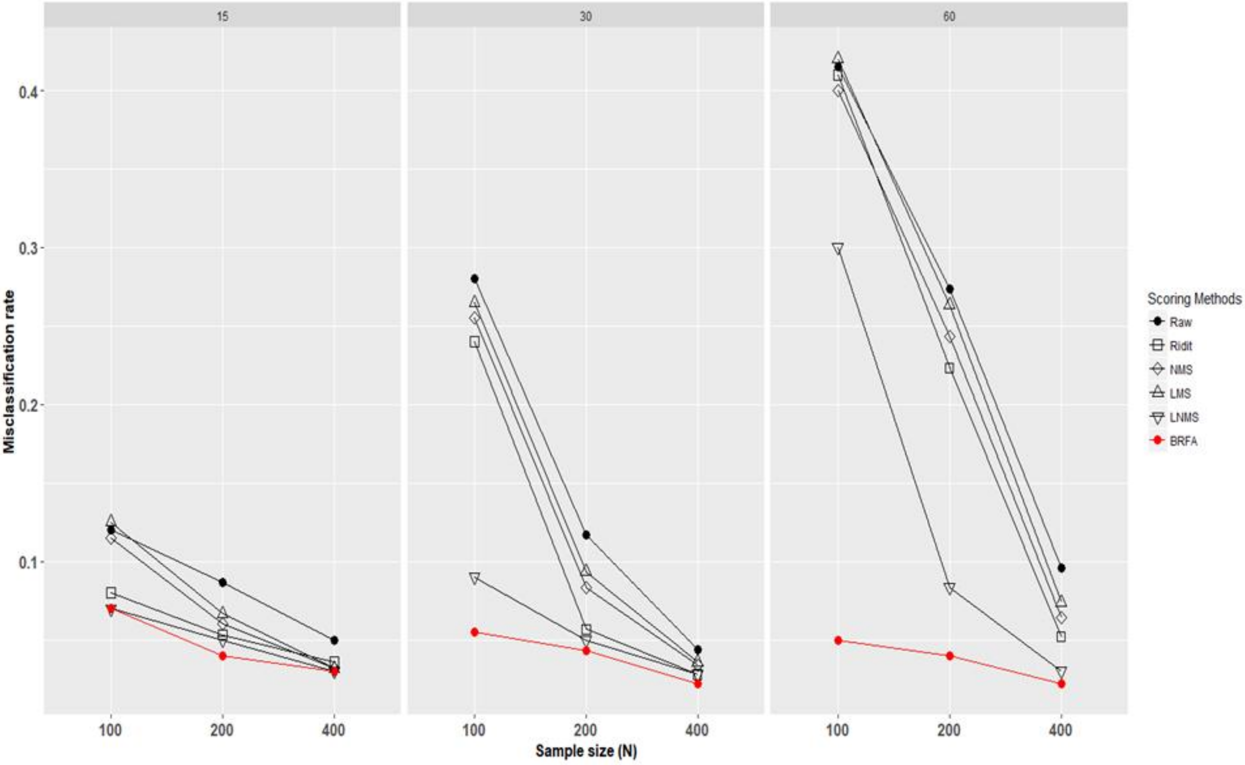

Figure A.29: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

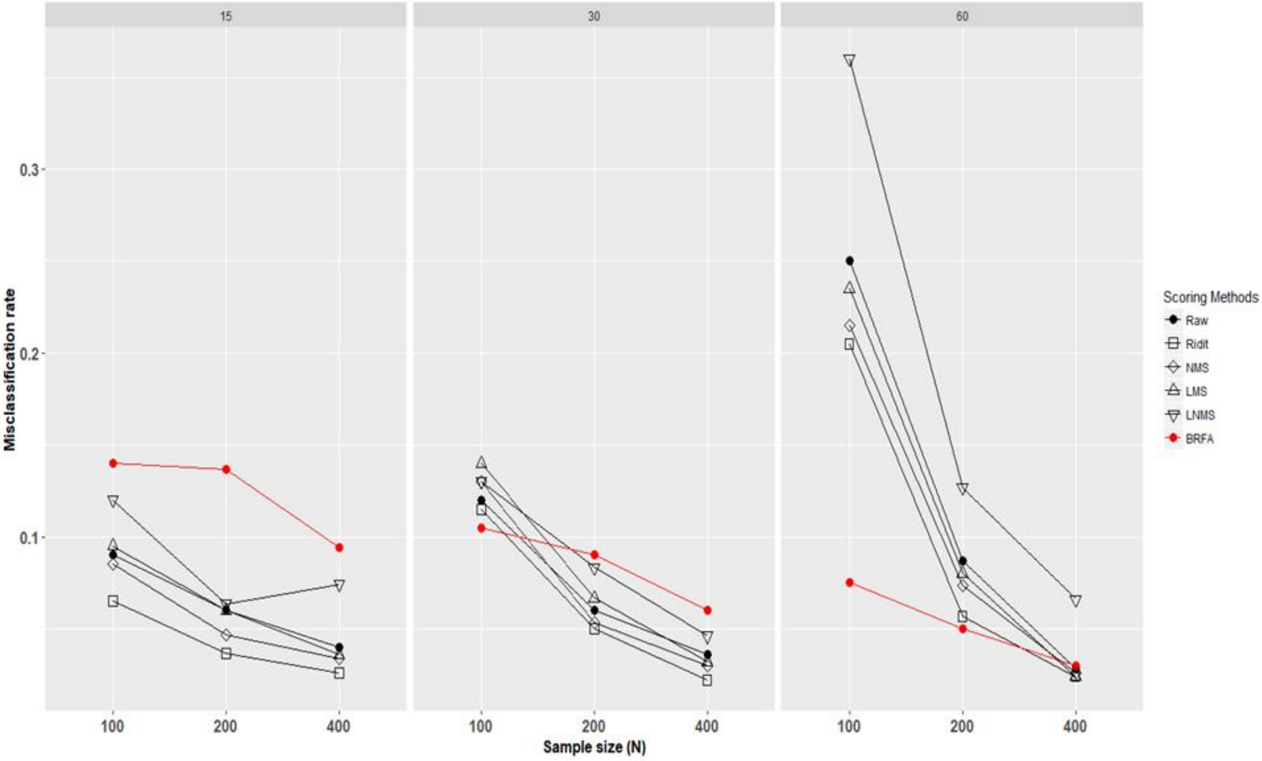

Figure A.30: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories $C = 357$.*

Table A.8: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for QDA in the four considered scenarios. Simulation results refer to the case of ordinal features with different number of categories ($C = 357$).*

	Scenario 1								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.35 (0.02)	0.29 (0.02)	0.22 (0.03)	0.42 (0.02)	0.32 (0.02)	0.24 (0.02)	0.39 (0.02)	0.42 (0.02)	0.33 (0.02)
Ridit	0.32 (0.04)	0.23 (0.02)	0.19 (0.02)	0.41 (0.01)	0.29 (0.02)	0.19 (0.01)	0.4 (0.01)	0.40 (0.02)	0.27 (0.02)
NM	0.35 (0.03)	0.28 (0.02)	0.21 (0.02)	0.43 (0.01)	0.32 (0.03)	0.24 (0.02)	0.41 (0.02)	0.42 (0.02)	0.34 (0.02)
Blom	0.35 (0.03)	0.28 (0.02)	0.21 (0.02)	0.43 (0.01)	0.32 (0.03)	0.24 (0.02)	0.41 (0.02)	0.42 (0.02)	0.34 (0.02)
NMS	0.34 (0.03)	0.28 (0.02)	0.21 (0.02)	0.43 (0.01)	0.32 (0.02)	0.24 (0.02)	0.40 (0.02)	0.41 (0.02)	0.33 (0.02)
LMS	0.34 (0.02)	0.30 (0.02)	0.22 (0.02)	0.42 (0.01)	0.33 (0.02)	0.25 (0.02)	0.40 (0.02)	0.43 (0.02)	0.34 (0.02)
LNMS	0.35 (0.03)	0.33 (0.01)	0.32 (0.02)	0.32 (0.03)	0.30 (0.02)	0.27 (0.02)	0.39 (0.02)	0.34 (0.02)	0.27 (0.02)
BRFA	0.22 (0.03)	0.25 (0.03)	0.17 (0.03)	0.13 (0.02)	0.12 (0.02)	0.08 (0.01)	0.07 (0.02)	0.08 (0.02)	0.06 (0.01)
	Scenario 2								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.15 (0.03)	0.13 (0.02)	0.06 (0.01)	0.23 (0.02)	0.10 (0.02)	0.04 (0.01)	0.38 (0.02)	0.28 (0.03)	0.11 (0.01)
Ridit	0.10 (0.02)	0.06 (0.01)	0.03 (0.01)	0.18 (0.03)	0.07 (0.01)	0.03 (0.01)	0.40 (0.02)	0.24 (0.02)	0.07 (0.01)
NM	0.10 (0.02)	0.10 (0.01)	0.04 (0.01)	0.22 (0.03)	0.09 (0.01)	0.04 (0.01)	0.38 (0.02)	0.27 (0.02)	0.10 (0.01)
Blom	0.10 (0.02)	0.11 (0.01)	0.04 (0.01)	0.22 (0.03)	0.09 (0.01)	0.04 (0.01)	0.38 (0.02)	0.27 (0.02)	0.10 (0.01)
NMS	0.10 (0.02)	0.08 (0.02)	0.04 (0.01)	0.18 (0.03)	0.09 (0.01)	0.04 (0.01)	0.40 (0.02)	0.26 (0.02)	0.10 (0.01)
LMS	0.12 (0.03)	0.11 (0.01)	0.04 (0.01)	0.22 (0.04)	0.09 (0.01)	0.04 (0.01)	0.41 (0.02)	0.27 (0.02)	0.10 (0.02)
LNMS	0.34 (0.02)	0.30 (0.03)	0.23 (0.03)	0.29 (0.03)	0.25 (0.04)	0.21 (0.02)	0.34 (0.03)	0.29 (0.03)	0.21 (0.02)
BRFA	0.13 (0.02)	0.12 (0.02)	0.07 (0.02)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
	Scenario 3								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.12 (0.02)	0.09 (0.01)	0.05 (0.01)	0.28 (0.02)	0.12 (0.02)	0.04 (0.01)	0.42 (0.02)	0.27 (0.03)	0.10 (0.01)
Ridit	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.24 (0.02)	0.06 (0.01)	0.03 (0.01)	0.41 (0.01)	0.22 (0.01)	0.05 (0.01)
NM	0.11 (0.02)	0.06 (0.02)	0.04 (0.01)	0.24 (0.03)	0.09 (0.01)	0.04 (0.01)	0.40 (0.02)	0.24 (0.03)	0.07 (0.01)
Blom	0.10 (0.02)	0.06 (0.02)	0.04 (0.01)	0.25 (0.02)	0.09 (0.01)	0.04 (0.01)	0.40 (0.02)	0.24 (0.03)	0.07 (0.01)
NMS	0.12 (0.02)	0.06 (0.02)	0.03 (0.01)	0.26 (0.02)	0.08 (0.01)	0.03 (0.01)	0.40 (0.02)	0.24 (0.02)	0.06 (0.01)
LMS	0.12 (0.02)	0.07 (0.02)	0.03 (0.01)	0.26 (0.03)	0.09 (0.01)	0.04 (0.01)	0.42 (0.03)	0.26 (0.02)	0.07 (0.01)
LNMS	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.09 (0.02)	0.05 (0.01)	0.03 (0.01)	0.30 (0.03)	0.08 (0.01)	0.03 (0.01)
BRFA	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)
	Mixture scenario								
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.09 (0.01)	0.06 (0.01)	0.04 (0.01)	0.12 (0.02)	0.06 (0.01)	0.04 (0.01)	0.25 (0.03)	0.09 (0.02)	0.03 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.12 (0.02)	0.05 (0.01)	0.02 (0.01)	0.2 (0.03)	0.06 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.12 (0.03)	0.05 (0.01)	0.03 (0.01)	0.2 (0.02)	0.07 (0.02)	0.03 (0.01)
Blom	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.12 (0.03)	0.05 (0.01)	0.03 (0.01)	0.2 (0.02)	0.07 (0.02)	0.03 (0.01)
NMS	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.13 (0.03)	0.05 (0.01)	0.03 (0.01)	0.22 (0.03)	0.07 (0.02)	0.03 (0.01)
LMS	0.1 (0.02)	0.06 (0.01)	0.04 (0.01)	0.14 (0.03)	0.07 (0.01)	0.03 (0.01)	0.24 (0.03)	0.08 (0.02)	0.02 (0.01)
LNMS	0.12 (0.01)	0.06 (0.01)	0.07 (0.01)	0.13 (0.02)	0.08 (0.02)	0.05 (0.01)	0.36 (0.03)	0.13 (0.02)	0.07 (0.01)
BRFA	0.14 (0.02)	0.14 (0.02)	0.09 (0.02)	0.10 (0.03)	0.09 (0.01)	0.06 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)

A.4 SVM

The support vector machine results to be one of the best classifier on these kind of data. In every scenario the results obtained through this classifier are almost always among the best ones, independently of the scoring methods. The scoring methods, in the majority of cases, do not significantly differ in term of the mean misclassification rates.

Generally, in the case $C = 5$ in scenario 2 there are not significant differences among the scoring methods, with the exception of LNMS (figure A.31). For scenarios 3 and "mixture" (figures A.32 and A.33) it is possible to notice that the BRFA misclassification rate is slightly higher than the others when $p = 10$ or $p = 20$ while, for higher values of p, it is the best or second best, although not significantly different.

As mentioned in chapter 5, the performance on LNMS greatly improve in scenario 3 (figure A.32). In this scenario the misclassification rate associated with LNMS is always the best or second best.

A similar situation is also observed in the case $C = 3$. From figures A.34 to A.37 it is possible to notice that in scenarios 1 and 2 the results associated with LNMS are generally worse than the other scoring methods, which have very similar mean misclassification rates. In "mixture" scenario for $p = 10$ and $p = 20$, as already observed for the other classifiers, it does not seem that BRFA is the optimal choice, while it is in absolute value the best, along with LNMS, for higher values of p. In scenario 3 the situation is similar to the case $C = 5$.

For $C = 357$ the behaviour of the misclassification rates associated to the scoring methods is similar to what observed previously, therefore we will not comment it further. It is to be noted, however, that although there are differences in the absolute value of mean misclassification rates, these are not significant considering the standard errors intervals reported in table A.11.

A.4.1 Simulation results for $C = 5$

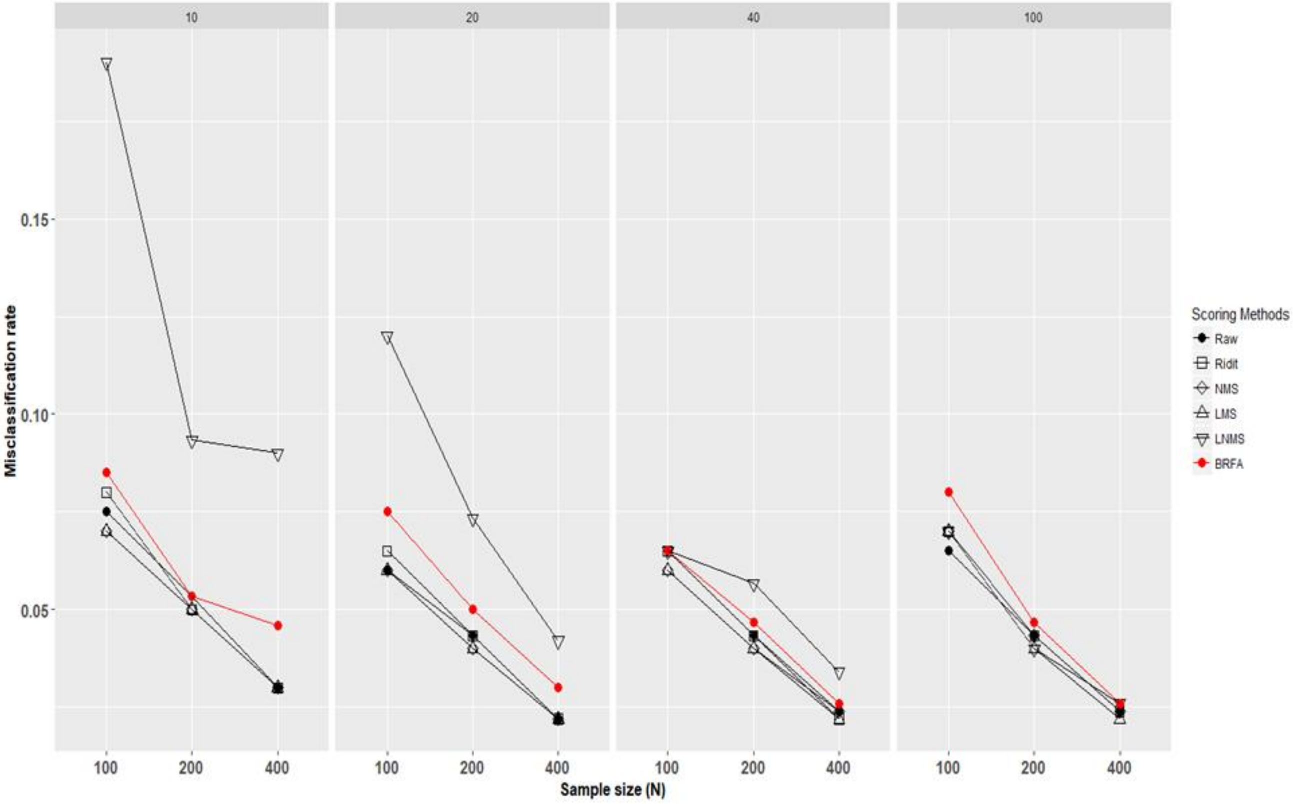

Figure A.31: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

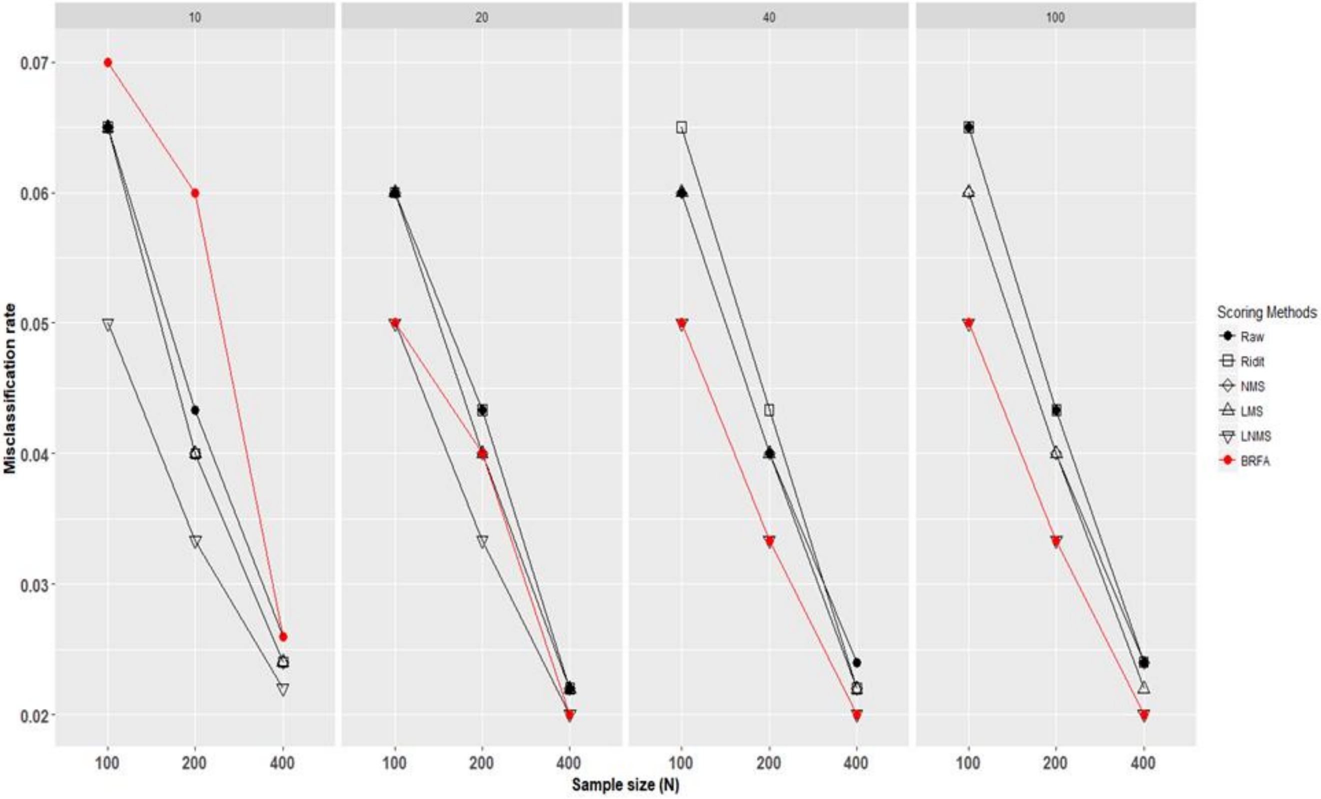

Figure A.32: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have* $C = 5$ *categories.*

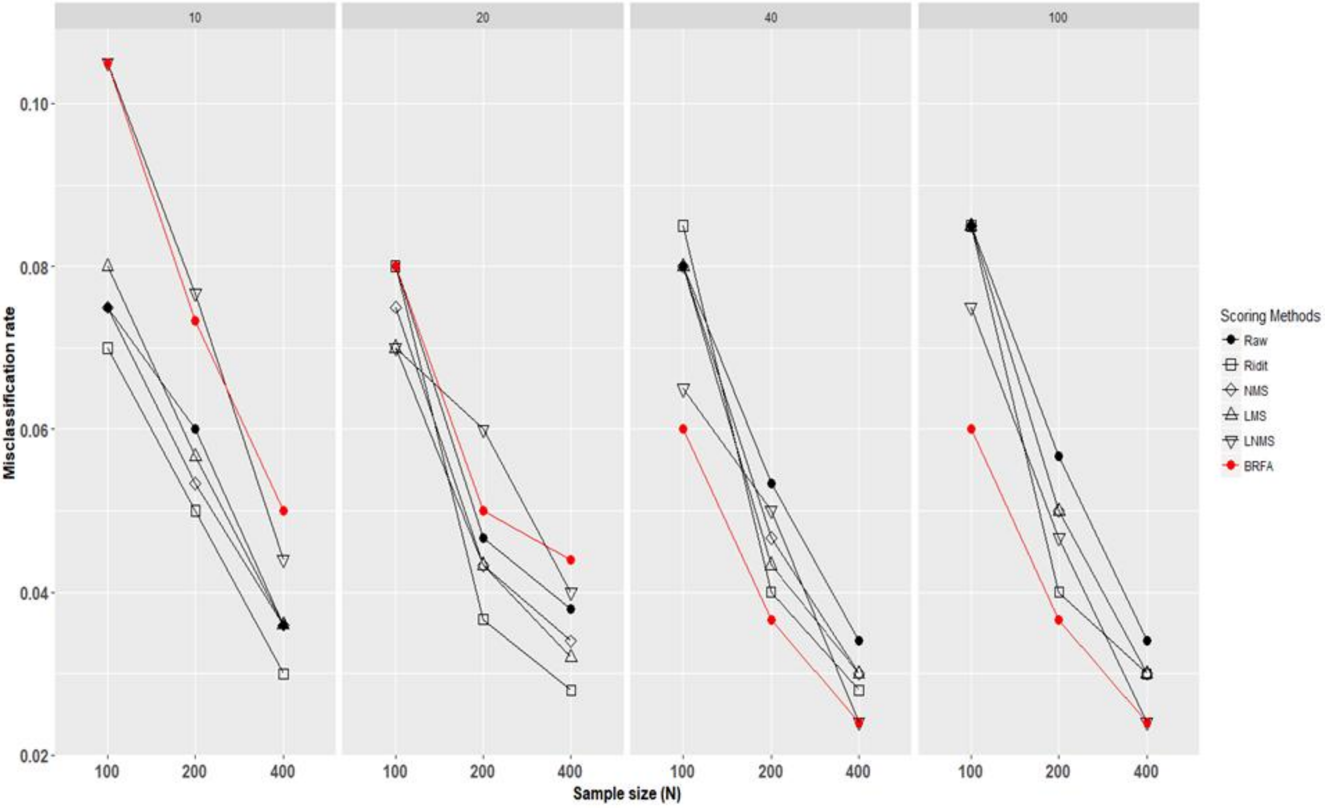

Figure A.33: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

Table A.9: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for SVM in the four considered scenarios. Simulation results refer to the case of ordinal features with five categories ($C = 5$).*

Scenario 1												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.20 (0.04)	0.19 (0.03)	0.19 (0.02)	0.17 (0.02)	0.14 (0.03)	0.11 (0.02)	0.1 (0.02)	0.07 (0.02)	0.06 (0.01)	0.10 (0.02)	0.06 (0.02)	0.05 (0.01)
Ridit	0.20 (0.03)	0.21 (0.02)	0.20 (0.02)	0.16 (0.02)	0.15 (0.03)	0.12 (0.02)	0.09 (0.02)	0.06 (0.02)	0.06 (0.01)	0.09 (0.02)	0.06 (0.01)	0.04 (0.01)
NM	0.19 (0.04)	0.20 (0.03)	0.20 (0.02)	0.17 (0.02)	0.14 (0.03)	0.11 (0.02)	0.09 (0.02)	0.07 (0.02)	0.06 (0.01)	0.10 (0.02)	0.06 (0.01)	0.05 (0.01)
Blom	0.20 (0.04)	0.20 (0.03)	0.20 (0.02)	0.17 (0.02)	0.14 (0.03)	0.11 (0.02)	0.09 (0.02)	0.07 (0.02)	0.06 (0.01)	0.10 (0.02)	0.06 (0.01)	0.05 (0.01)
NMS	0.20 (0.04)	0.20 (0.03)	0.20 (0.02)	0.17 (0.02)	0.14 (0.03)	0.11 (0.02)	0.09 (0.02)	0.07 (0.02)	0.06 (0.01)	0.10 (0.02)	0.06 (0.01)	0.05 (0.01)
LMS	0.20 (0.04)	0.20 (0.03)	0.20 (0.02)	0.16 (0.02)	0.13 (0.02)	0.11 (0.02)	0.09 (0.02)	0.07 (0.02)	0.06 (0.01)	0.10 (0.02)	0.06 (0.02)	0.05 (0.01)
LNMS	0.28 (0.03)	0.22 (0.03)	0.21 (0.02)	0.20 (0.03)	0.18 (0.03)	0.17 (0.02)	0.14 (0.02)	0.16 (0.02)	0.13 (0.01)	0.14 (0.03)	0.12 (0.02)	0.13 (0.01)
BRFA	0.18 (0.03)	0.21 (0.03)	0.20 (0.01)	0.18 (0.02)	0.17 (0.02)	0.15 (0.01)	0.08 (0.02)	0.07 (0.02)	0.06 (0.01)	0.08 (0.01)	0.06 (0.01)	0.04 (0.01)
Scenario 2												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Ridit	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
LMS	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)
LNMS	0.19 (0.03)	0.09 (0.02)	0.09 (0.01)	0.12 (0.02)	0.07 (0.01)	0.04 (0.01)	0.06 (0.02)	0.06 (0.01)	0.03 (0.01)	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)
BRFA	0.08 (0.02)	0.05 (0.01)	0.05 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
Scenario 3												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
LNMS	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.07 (0.02)	0.06 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Mixture scenario												
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.06 (0.02)	0.04 (0.01)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.08 (0.02)	0.06 (0.01)	0.03 (0.01)
Ridit	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)
NM	0.08 (0.02)	0.05 (0.02)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.06 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
Blom	0.08 (0.02)	0.05 (0.02)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.06 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
NMS	0.08 (0.02)	0.05 (0.02)	0.04 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
LMS	0.08 (0.02)	0.06 (0.02)	0.04 (0.01)	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)
LNMS	0.10 (0.02)	0.08 (0.01)	0.04 (0.01)	0.07 (0.01)	0.06 (0.01)	0.04 (0.01)	0.06 (0.01)	0.05 (0.01)	0.02 (0.01)	0.08 (0.02)	0.05 (0.01)	0.02 (0.01)
BRFA	0.10 (0.02)	0.07 (0.01)	0.05 (0.01)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)

A.4.2 Simulation results for $C = 3$

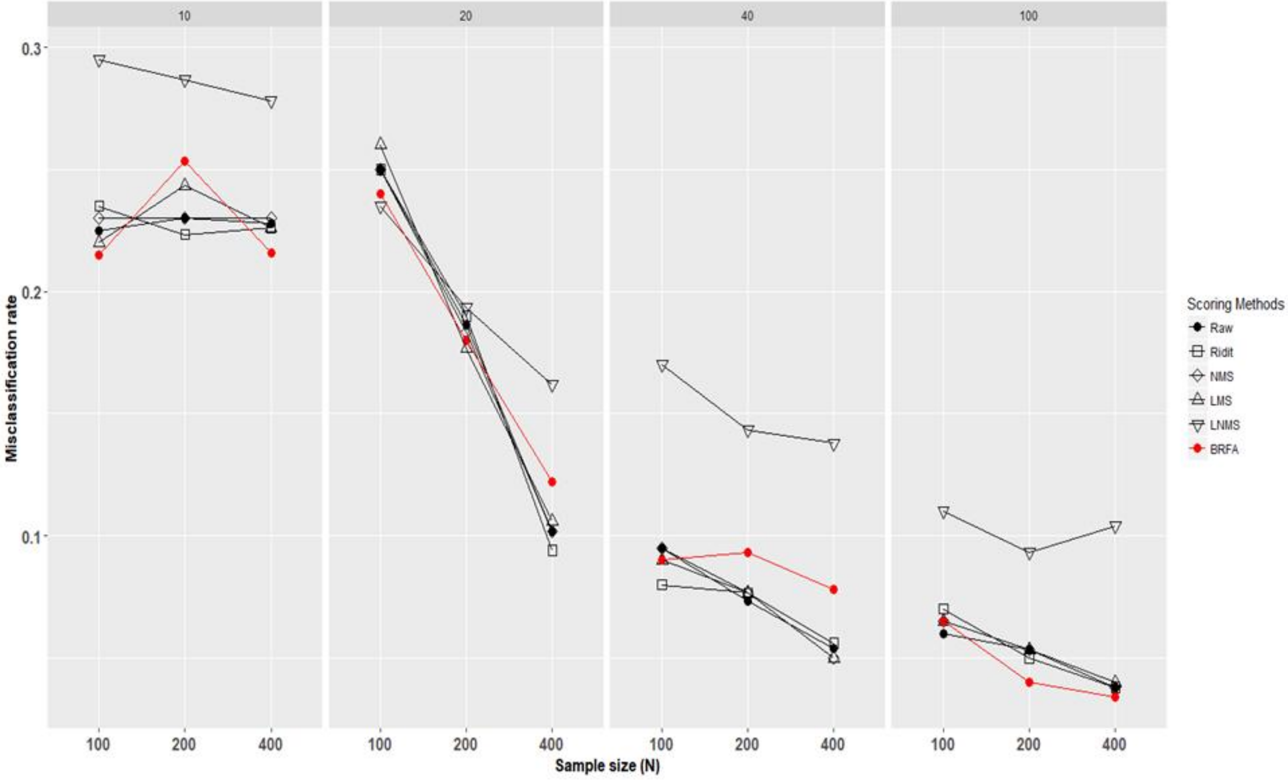

Figure A.34: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

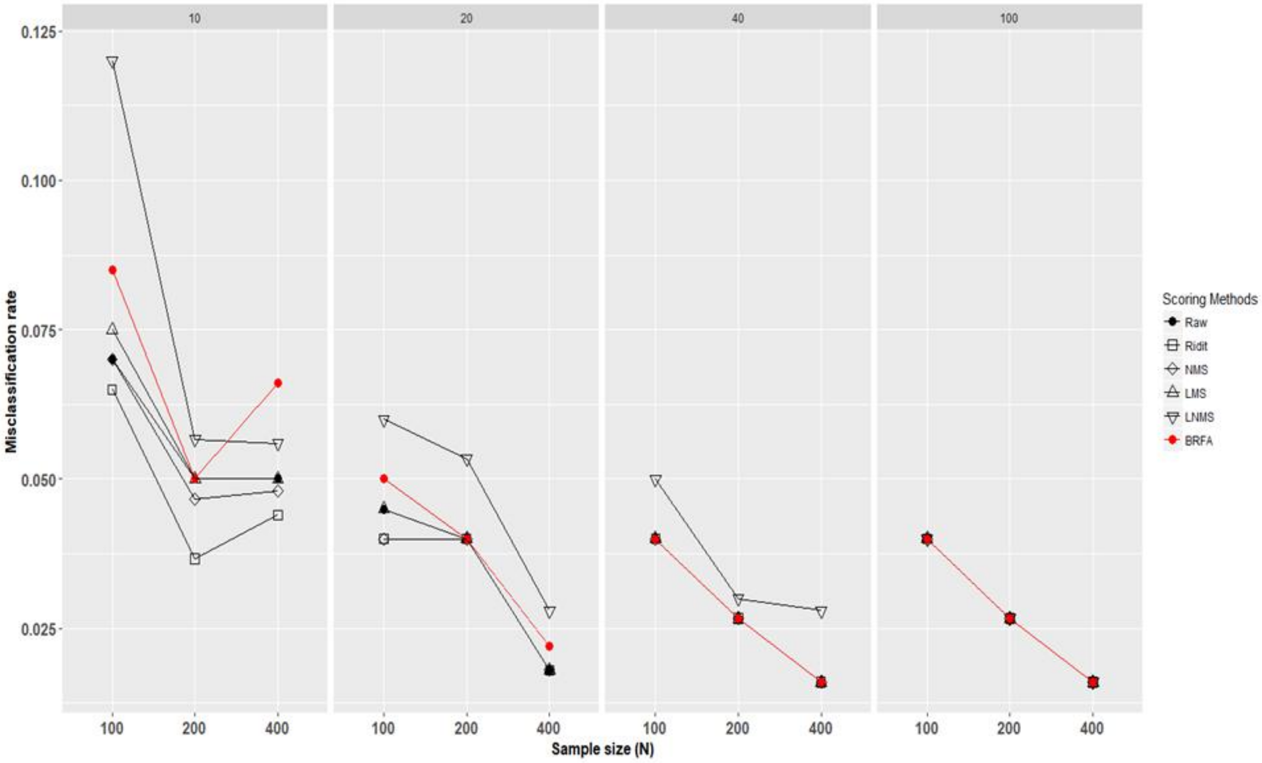

Figure A.35: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

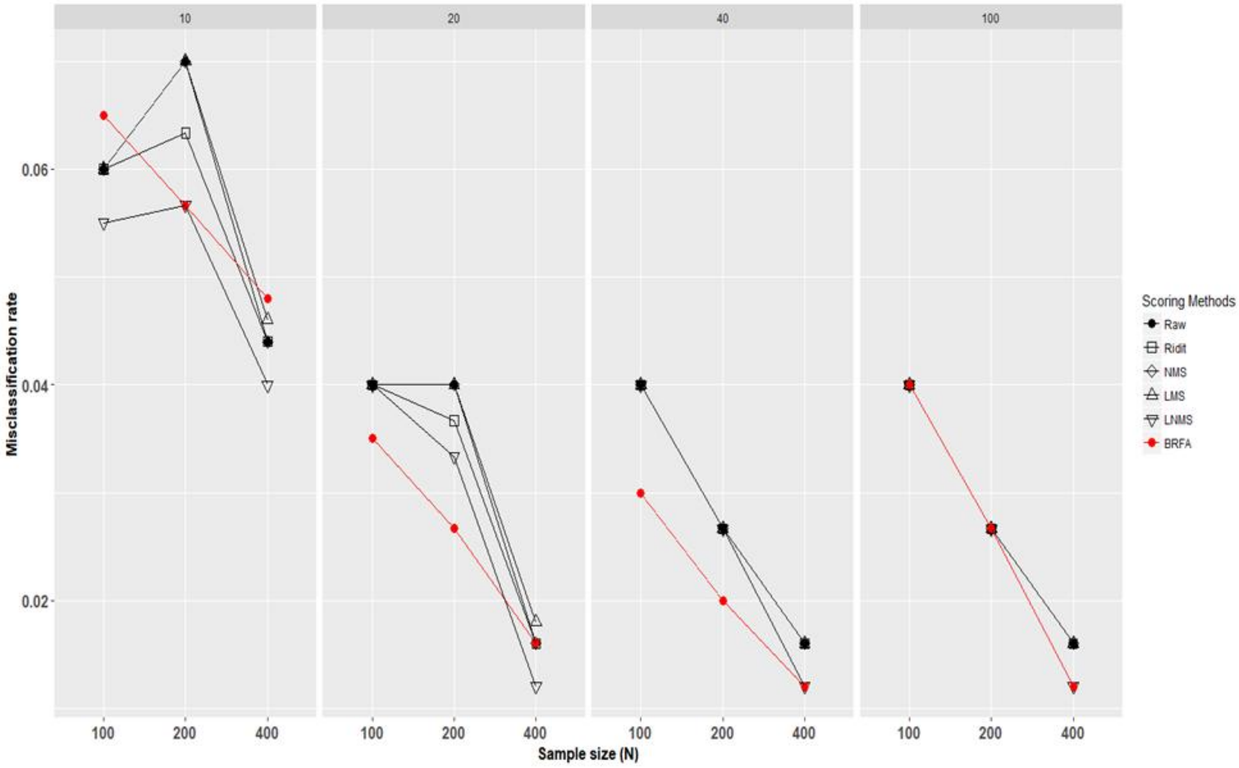

Figure A.36: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

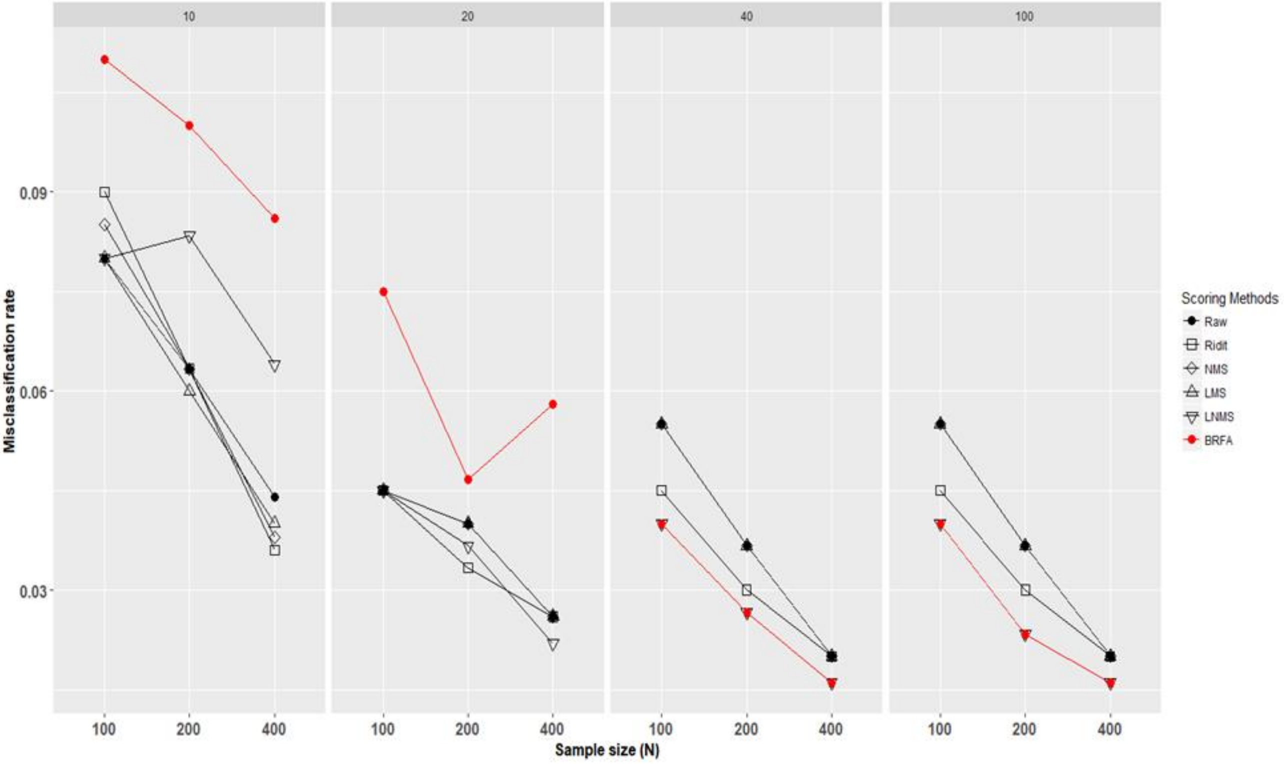

Figure A.37: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

Table A.10: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for SVM in the four considered scenarios. Simulation results refer to the case of ordinal features with three categories ($C = 3$).*

	Scenario 1											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.22 (0.03)	0.23 (0.02)	0.23 (0.01)	0.25 (0.02)	0.19 (0.01)	0.10 (0.01)	0.1 (0.02)	0.07 (0.02)	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)
Ridit	0.24 (0.03)	0.22 (0.02)	0.23 (0.02)	0.25 (0.02)	0.19 (0.01)	0.09 (0.01)	0.08 (0.02)	0.08 (0.01)	0.06 (0.01)	0.07 (0.01)	0.05 (0.01)	0.04 (0.01)
NM	0.22 (0.03)	0.24 (0.03)	0.23 (0.02)	0.26 (0.02)	0.18 (0.02)	0.11 (0.01)	0.10 (0.02)	0.08 (0.01)	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)
Blom	0.22 (0.03)	0.24 (0.03)	0.23 (0.02)	0.26 (0.02)	0.18 (0.02)	0.11 (0.01)	0.10 (0.02)	0.08 (0.01)	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)
NMS	0.23 (0.03)	0.23 (0.02)	0.23 (0.01)	0.25 (0.02)	0.18 (0.02)	0.10 (0.01)	0.10 (0.02)	0.08 (0.01)	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)
LMS	0.22 (0.03)	0.24 (0.02)	0.23 (0.02)	0.26 (0.02)	0.18 (0.02)	0.11 (0.01)	0.09 (0.02)	0.08 (0.01)	0.05 (0.01)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)
LNMS	0.30 (0.03)	0.29 (0.03)	0.28 (0.02)	0.24 (0.01)	0.19 (0.03)	0.16 (0.02)	0.17 (0.02)	0.14 (0.02)	0.14 (0.01)	0.11 (0.03)	0.09 (0.01)	0.10 (0.01)
BRFA	0.22 (0.03)	0.25 (0.02)	0.22 (0.01)	0.24 (0.03)	0.18 (0.02)	0.12 (0.01)	0.09 (0.02)	0.09 (0.02)	0.08 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)
	Scenario 2											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.02)	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.07 (0.02)	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
Blom	0.07 (0.02)	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
NMS	0.07 (0.02)	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
LMS	0.08 (0.02)	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
LNMS	0.12 (0.02)	0.06 (0.02)	0.06 (0.01)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
BRFA	0.08 (0.02)	0.05 (0.01)	0.07 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
	Scenario 3											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.01)	0.07 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
Ridit	0.06 (0.01)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
NM	0.06 (0.01)	0.07 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
Blom	0.06 (0.01)	0.07 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
NMS	0.06 (0.01)	0.07 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
LMS	0.06 (0.01)	0.07 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
LNMS	0.06 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.01 (0.01)	0.04 (0.01)	0.03 (0.01)	0.01 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
BRFA	0.06 (0.02)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)
	Mixture scenario											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
Ridit	0.09 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
Blom	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
NMS	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
LMS	0.08 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)
LNMS	0.08 (0.02)	0.08 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)
BRFA	0.11 (0.02)	0.10 (0.01)	0.09 (0.01)	0.08 (0.01)	0.05 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)

A.4.3 Simulation results for $C = 357$

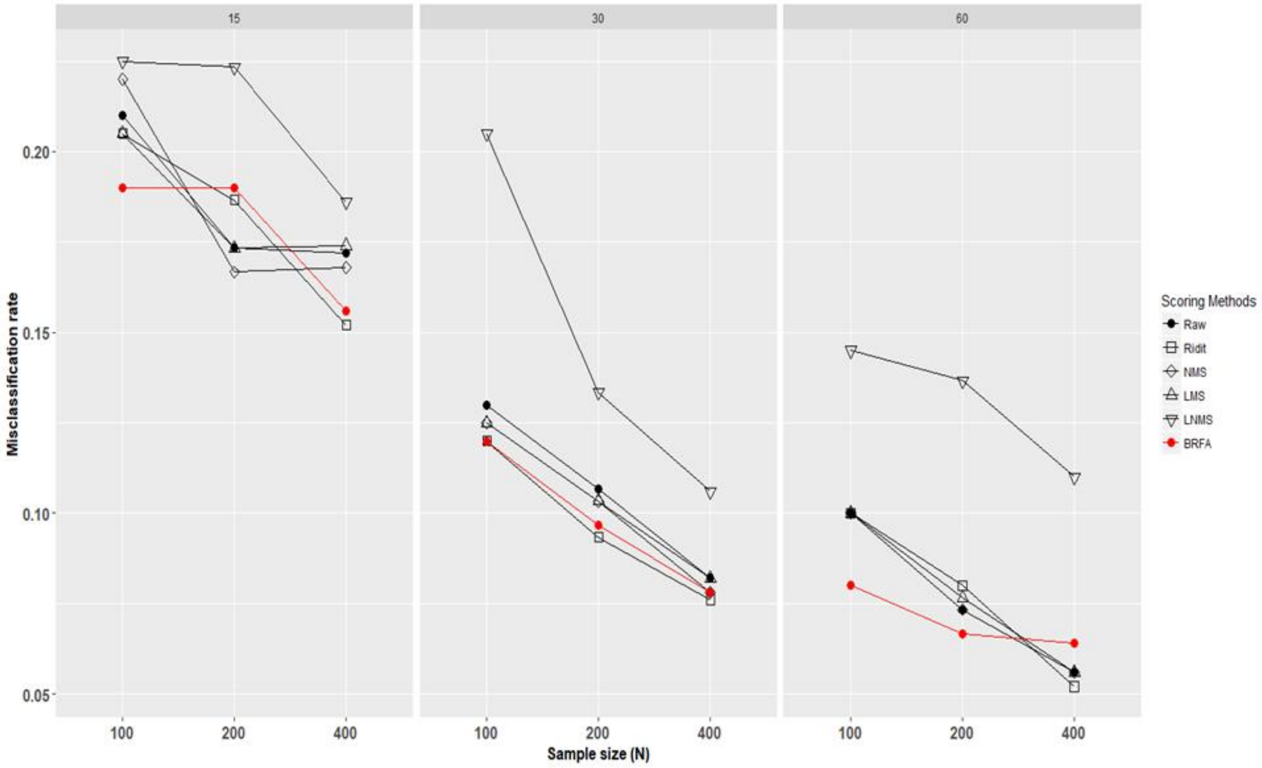

Figure A.38: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

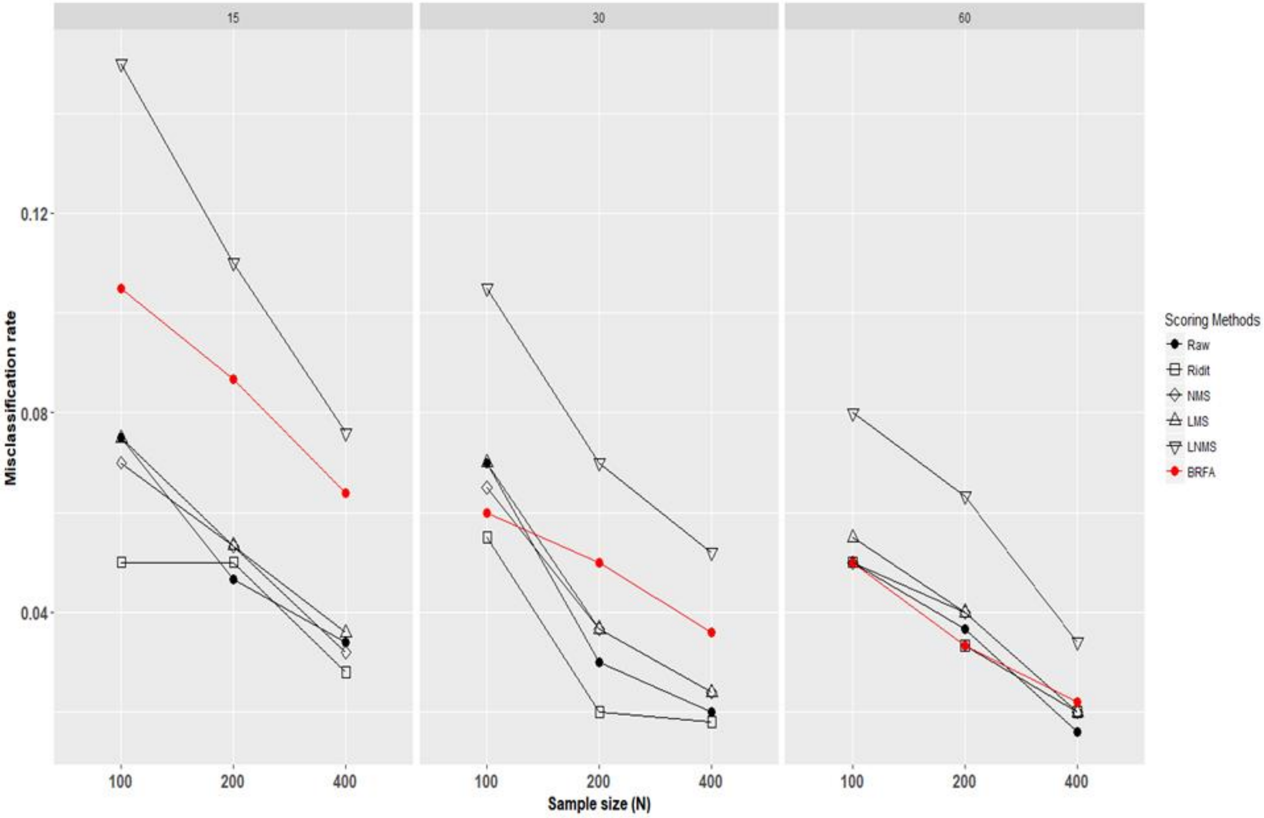

Figure A.39: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

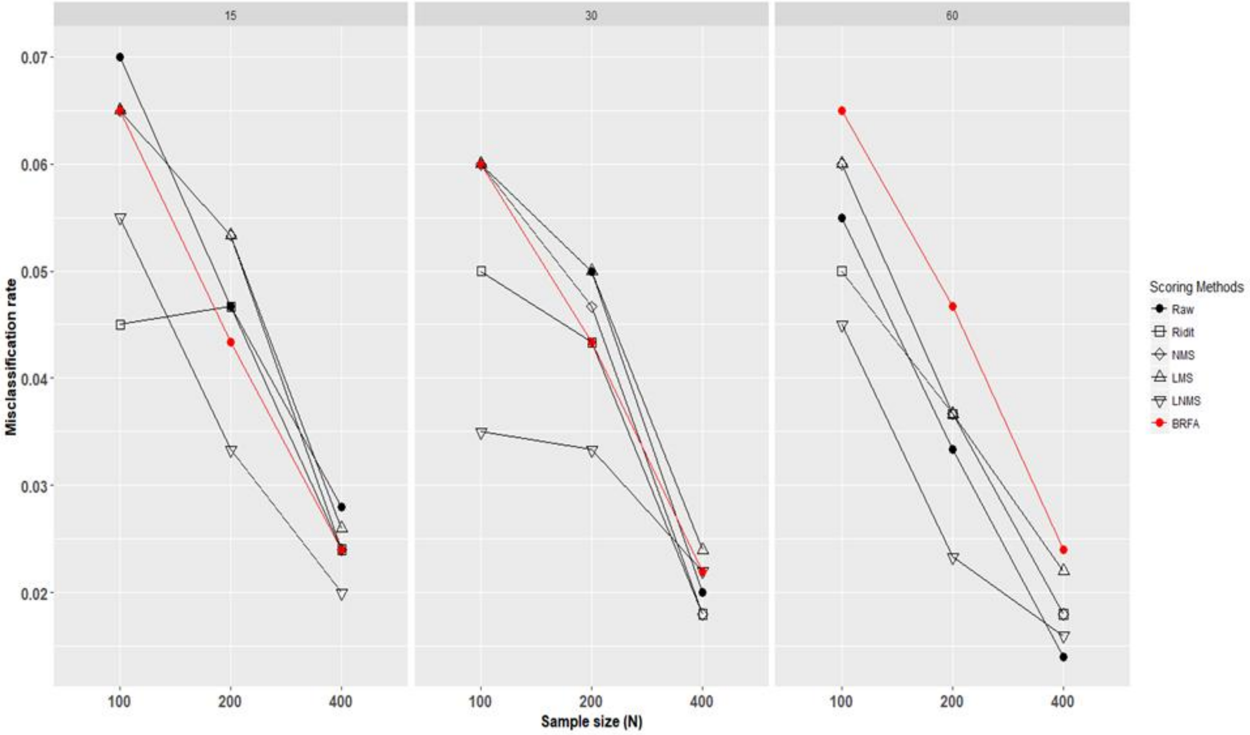

Figure A.40: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories $C = 357$.*

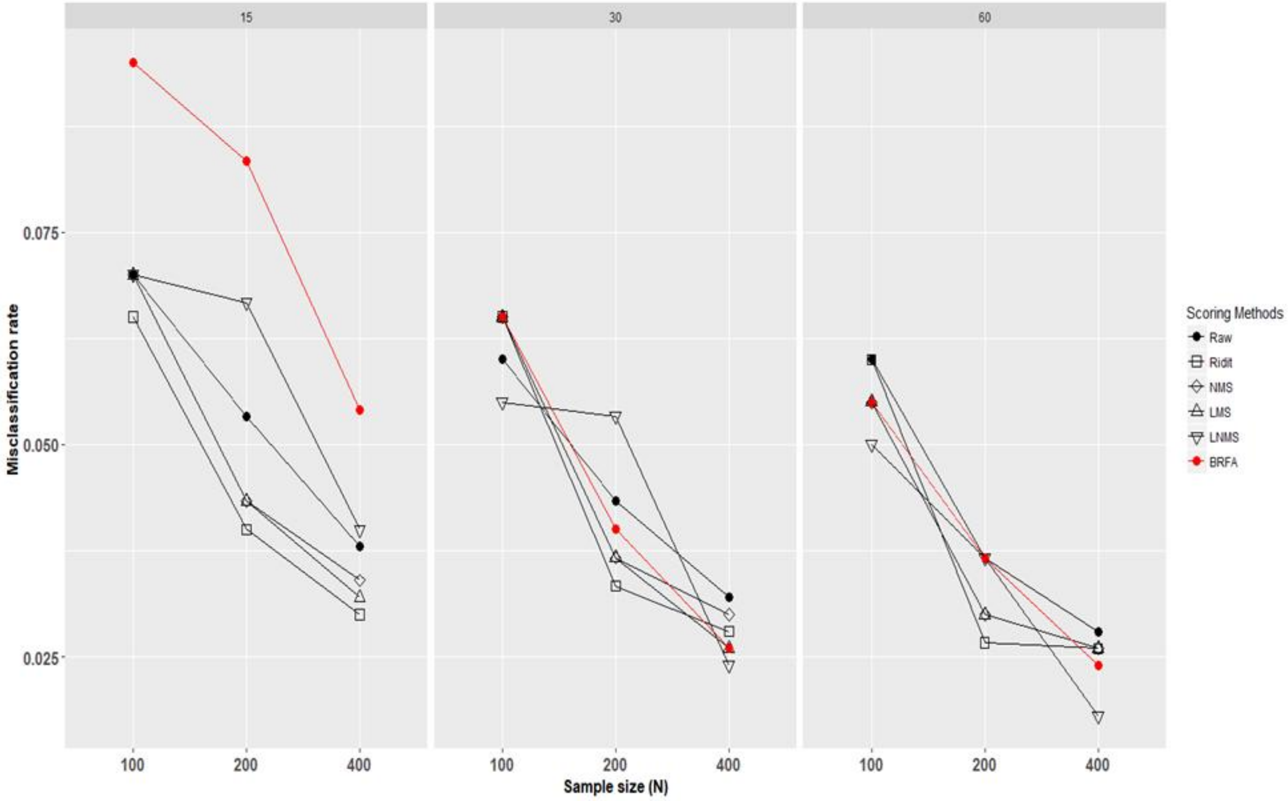

Figure A.41: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories $C = 357$.*

Table A.11: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for SVM in the four considered scenarios. Simulation results refer to the case of ordinal features with different number of categories ($C = 357$).*

					Scenario 1				
		p=15			**p=30**			**p=60**	
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.21 (0.03)	0.17 (0.01)	0.17 (0.01)	0.13 (0.02)	0.11 (0.02)	0.08 (0.01)	0.10 (0.03)	0.07 (0.02)	0.06 (0.01)
Ridit	0.2 (0.03)	0.19 (0.01)	0.15 (0.01)	0.12 (0.02)	0.09 (0.02)	0.08 (0.02)	0.10 (0.03)	0.08 (0.02)	0.05 (0.01)
NM	0.22 (0.03)	0.17 (0.01)	0.17 (0.01)	0.12 (0.02)	0.10 (0.02)	0.08 (0.01)	0.10 (0.03)	0.07 (0.02)	0.06 (0.01)
Blom	0.22 (0.03)	0.17 (0.01)	0.17 (0.01)	0.12 (0.02)	0.10 (0.02)	0.08 (0.01)	0.10 (0.03)	0.07 (0.02)	0.06 (0.01)
NMS	0.22 (0.03)	0.17 (0.01)	0.17 (0.01)	0.12 (0.02)	0.10 (0.02)	0.08 (0.02)	0.10 (0.03)	0.07 (0.02)	0.06 (0.01)
LMS	0.20 (0.03)	0.17 (0.01)	0.17 (0.02)	0.12 (0.02)	0.10 (0.02)	0.08 (0.02)	0.10 (0.03)	0.08 (0.02)	0.06 (0.01)
LNMS	0.22 (0.03)	0.22 (0.01)	0.19 (0.02)	0.20 (0.02)	0.13 (0.03)	0.11 (0.01)	0.14 (0.02)	0.14 (0.02)	0.11 (0.02)
BRFA	0.19 (0.03)	0.19 (0.01)	0.16 (0.02)	0.12 (0.02)	0.10 (0.01)	0.08 (0.01)	0.08 (0.02)	0.07 (0.01)	0.06 (0.01)
					Scenario 2				
		p=15			**p=30**			**p=60**	
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)	0.07 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
Ridit	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.02 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
NM	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
LMS	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.07 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
LNMS	0.15 (0.03)	0.11 (0.02)	0.08 (0.01)	0.10 (0.02)	0.07 (0.01)	0.05 (0.01)	0.08 (0.02)	0.06 (0.02)	0.03 (0.01)
BRFA	0.10 (0.02)	0.09 (0.02)	0.06 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
					Scenario 3				
		p=15			**p=30**			**p=60**	
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.01 (0.01)
Ridit	0.04 (0.01)	0.05 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
LNMS	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.04 (0.01)	0.02 (0.01)	0.02 (0.01)
BRFA	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.05 (0.01)	0.02 (0.01)
					Mixture scenario				
		p=15			**p=30**			**p=60**	
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.01)	0.05 (0.02)	0.04 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)
Ridit	0.06 (0.01)	0.04 (0.02)	0.03 (0.01)	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)
NM	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)
Blom	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)
NMS	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)
LMS	0.07 (0.02)	0.04 (0.02)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)
LNMS	0.07 (0.01)	0.07 (0.01)	0.04 (0.01)	0.06 (0.01)	0.05 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)
BRFA	0.10 (0.02)	0.08 (0.02)	0.05 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)

A.5 Naive Bayes

For $C = 5$ the mean misclassification rates obtained by applying the naive Bayes classifier to the data scored through BRFA are generally the highest ones in scenario 2 and “mixture” (figures A.42 and A.44), excluding the log-normal scores. Our method of scoring does not seem to be the most appropriate for the naive Bayes classifier which, as mentioned in chapter 5, is the classifier that performs best (together with SVM) on the simulated datasets. By reducing the p ordinal variables to a single continuous variable with BRFA the naive Bayes classifier provides generally worse results in the simulations. The scenario 3 (figure A.43) is an exception as, for $p = 40$ and $p = 100$, the scoring methods applied to highly skewed and well-separated data provide identical results. Generally, it is possible to notice that, except for scenario 2, the mean misclassification rates among the scoring methods are not significantly different in high dimensional datasets.

Also for the naive Bayes classifier the log-normal scores seems to not be appropriate for these kind of data (with the exception of scenario 3) and are associated with higher misclassification rates.

As for the other classifiers, the naive Bayes behaves in a similar way for $C = 5$ and $C = 3$. For this reason the latter case will not be further commented.

In the case of features with mixed number of possible categories (figures A.49 to A.52) in scenario 1, 2 and “mixture” for high dimensional datasets the scoring methods do not significantly differ while, for lower values of p, the BRFA results are generally worse. In scenario 3 there are not significant differences among scoring methods for every value of p.

A.5.1 Simulation results for $C = 5$

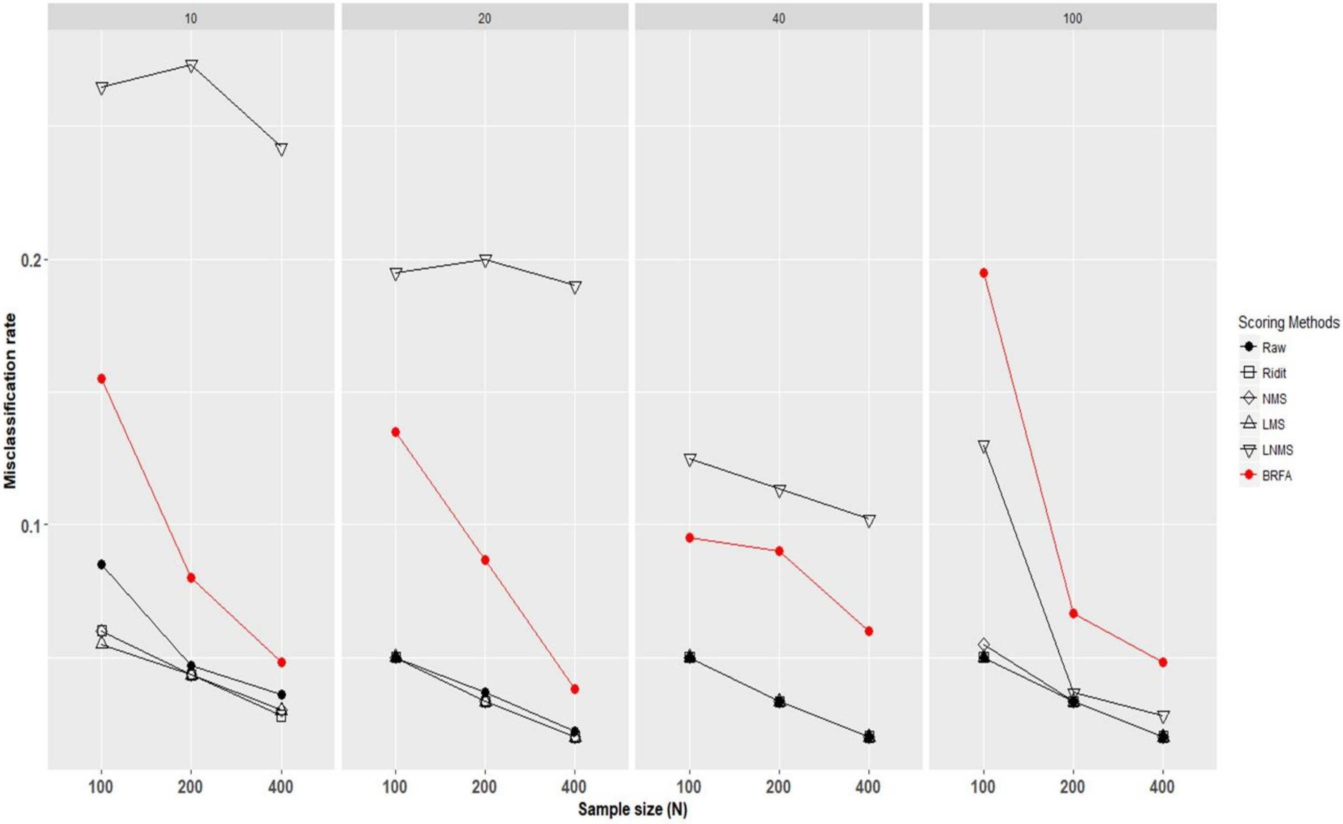

Figure A.42: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

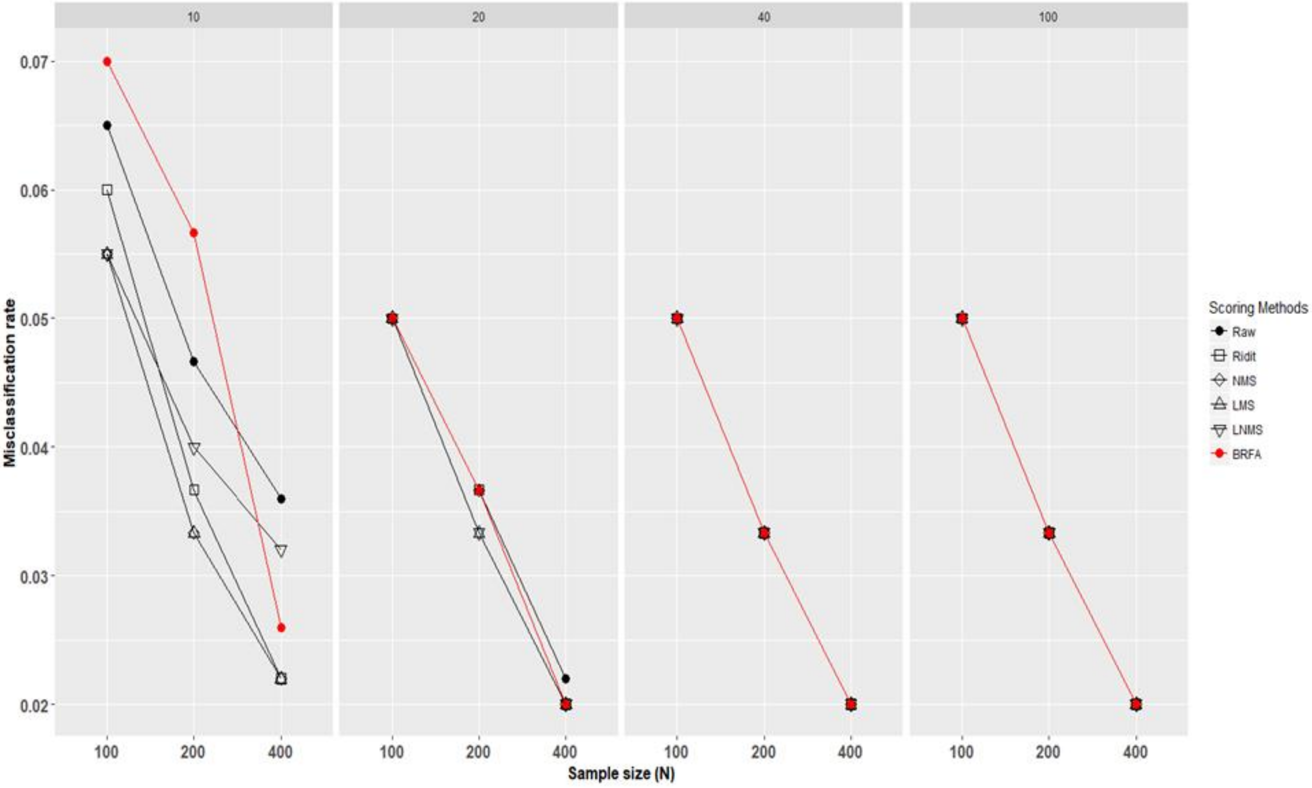

Figure A.43: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

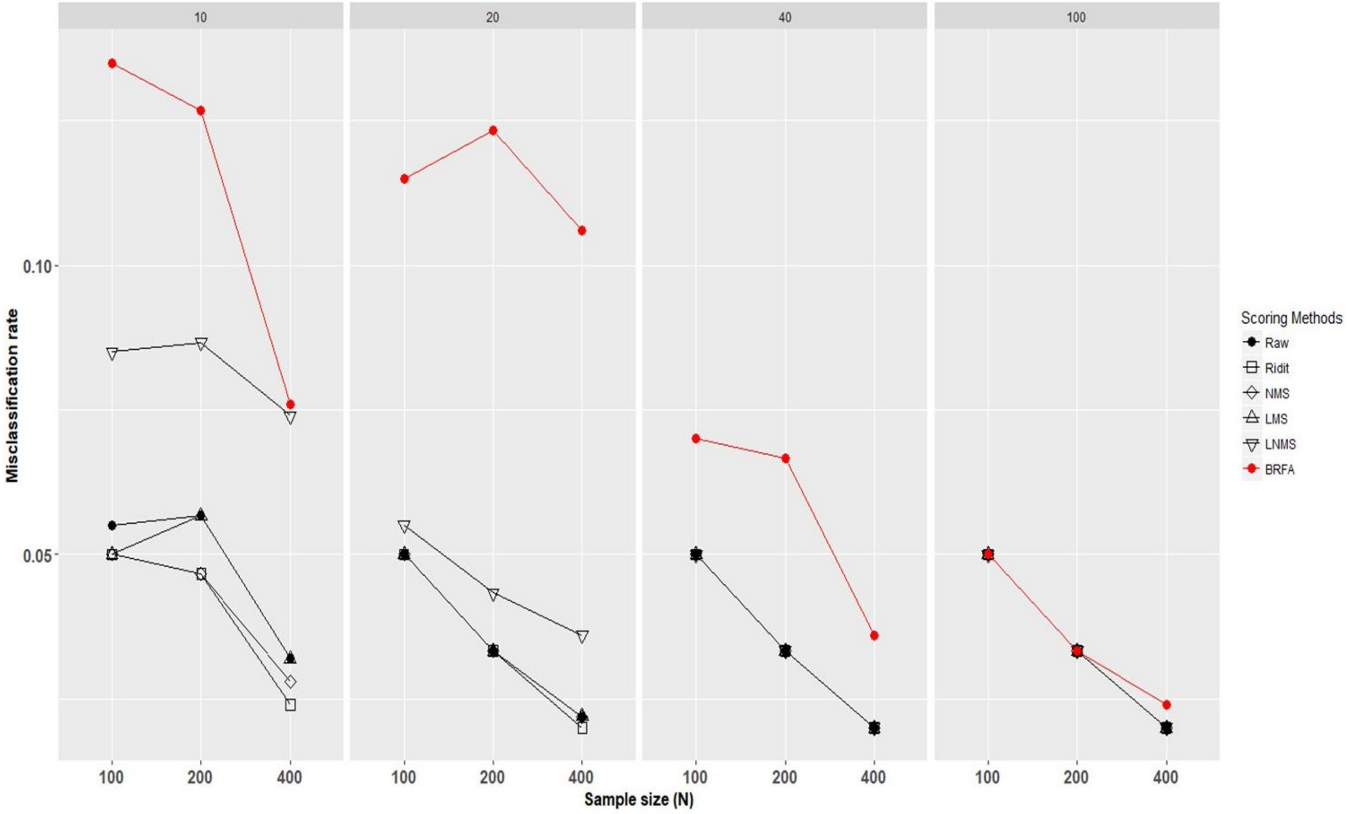

Figure A.44: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 5$ categories.*

Table A.12: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for naive Bayes in the four considered scenarios. Simulation results refer to the case of ordinal features with five categories ($C = 5$).*

						Scenario 1						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.22 (0.03)	0.2 (0.03)	0.19 (0.02)	0.17 (0.02)	0.13 (0.03)	0.11 (0.02)	0.06 (0.02)	0.06 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)
Ridit	0.20 (0.04)	0.17 (0.02)	0.19 (0.02)	0.14 (0.02)	0.13 (0.02)	0.10 (0.02)	0.06 (0.02)	0.05 (0.02)	0.05 (0.01)	0.08 (0.02)	0.04 (0.01)	0.04 (0.01)
NM	0.20 (0.03)	0.17 (0.02)	0.19 (0.02)	0.14 (0.01)	0.13 (0.02)	0.10 (0.02)	0.06 (0.01)	0.05 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)
Blom	0.20 (0.03)	0.17 (0.02)	0.19 (0.02)	0.14 (0.01)	0.13 (0.02)	0.10 (0.02)	0.06 (0.01)	0.05 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)
NMS	0.19 (0.03)	0.17 (0.02)	0.19 (0.02)	0.14 (0.01)	0.13 (0.02)	0.10 (0.02)	0.06 (0.01)	0.05 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)
LMS	0.20 (0.03)	0.18 (0.02)	0.19 (0.02)	0.16 (0.02)	0.13 (0.02)	0.10 (0.02)	0.06 (0.01)	0.06 (0.02)	0.05 (0.01)	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)
LNMS	0.34 (0.03)	0.35 (0.02)	0.38 (0.01)	0.36 (0.03)	0.36 (0.02)	0.37 (0.02)	0.27 (0.02)	0.36 (0.02)	0.35 (0.02)	0.32 (0.02)	0.30 (0.03)	0.31 (0.03)
BRFA	0.22 (0.03)	0.24 (0.03)	0.23 (0.03)	0.20 (0.03)	0.22 (0.03)	0.15 (0.01)	0.06 (0.01)	0.08 (0.02)	0.07 (0.01)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)
						Scenario 2						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Ridit	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Blom	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NMS	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)
LMS	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LNMS	0.26 (0.04)	0.27 (0.03)	0.24 (0.01)	0.20 (0.03)	0.2 (0.01)	0.19 (0.02)	0.12 (0.02)	0.11 (0.02)	0.10 (0.02)	0.13 (0.02)	0.04 (0.01)	0.03 (0.01)
BRFA	0.16 (0.03)	0.08 (0.01)	0.05 (0.01)	0.14 (0.03)	0.09 (0.02)	0.04 (0.01)	0.10 (0.02)	0.09 (0.01)	0.06 (0.01)	0.20 (0.04)	0.07 (0.01)	0.05 (0.01)
						Scenario 3						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
LNMS	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.07 (0.02)	0.06 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
						Mixture scenario						
		p=10			p=20			p=40			p=100	
	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400	N=100	N=200	N=400
Raw	0.06 (0.02)	0.06 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
Ridit	0.05 (0.02)	0.05 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
NM	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
Blom	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
NMS	0.05 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LMS	0.05 (0.02)	0.06 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LNMS	0.08 (0.02)	0.09 (0.01)	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)
BRFA	0.14 (0.02)	0.13 (0.02)	0.08 (0.01)	0.12 (0.03)	0.12 (0.02)	0.11 (0.01)	0.07 (0.02)	0.07 (0.01)	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)

A.5.2 Simulation results for $C = 3$

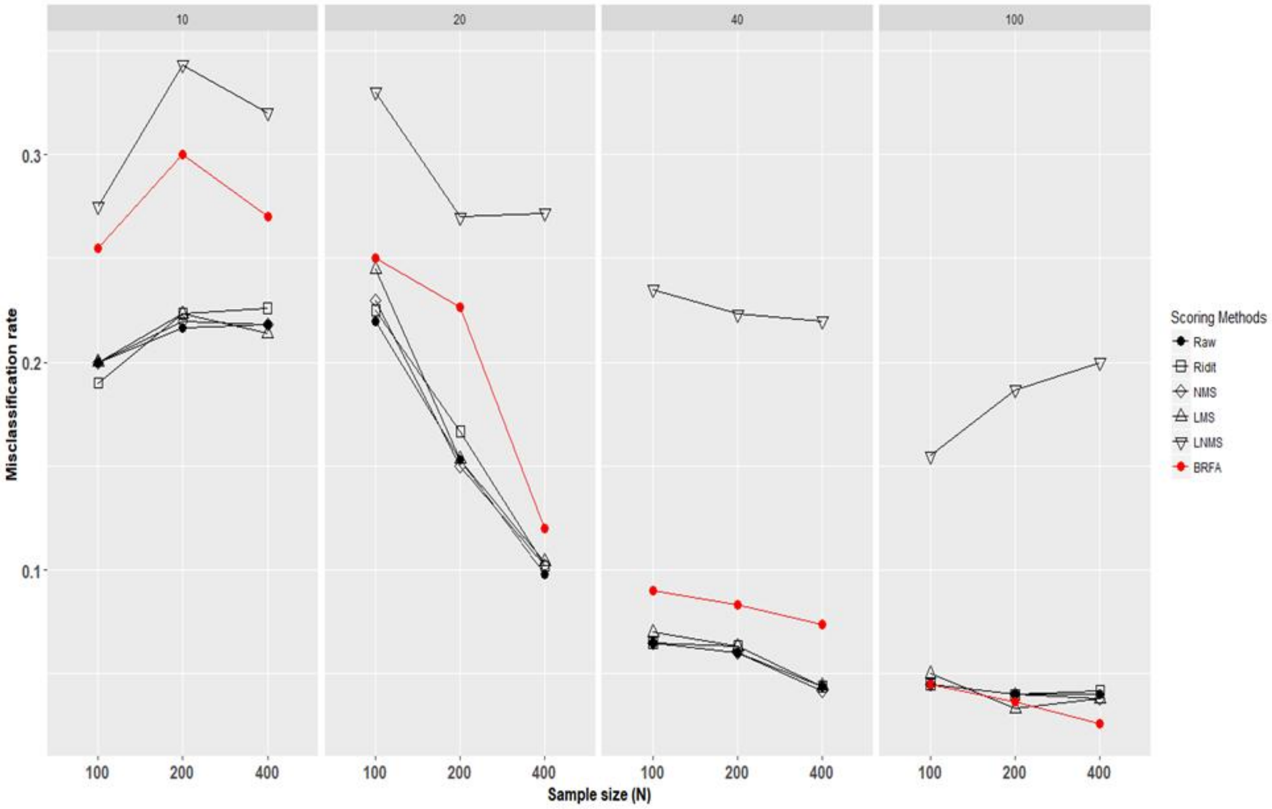

Figure A.45: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

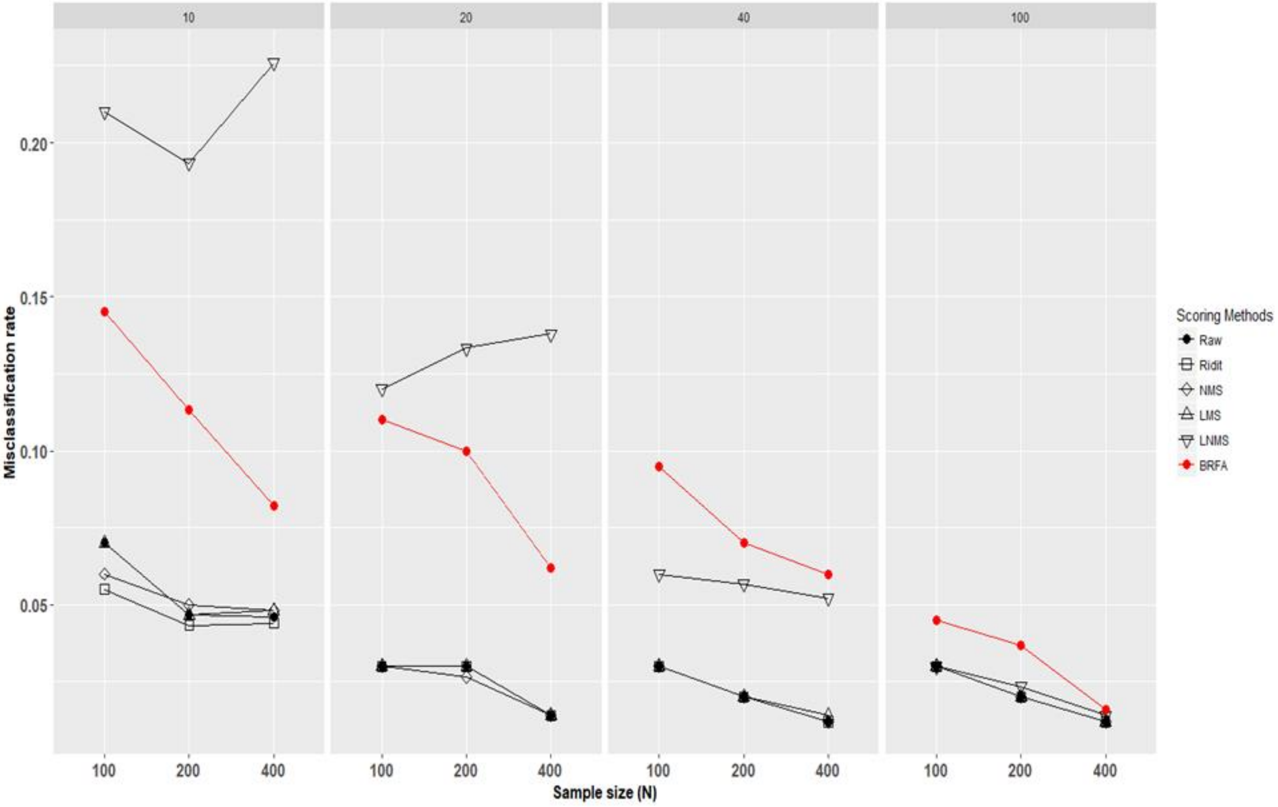

Figure A.46: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

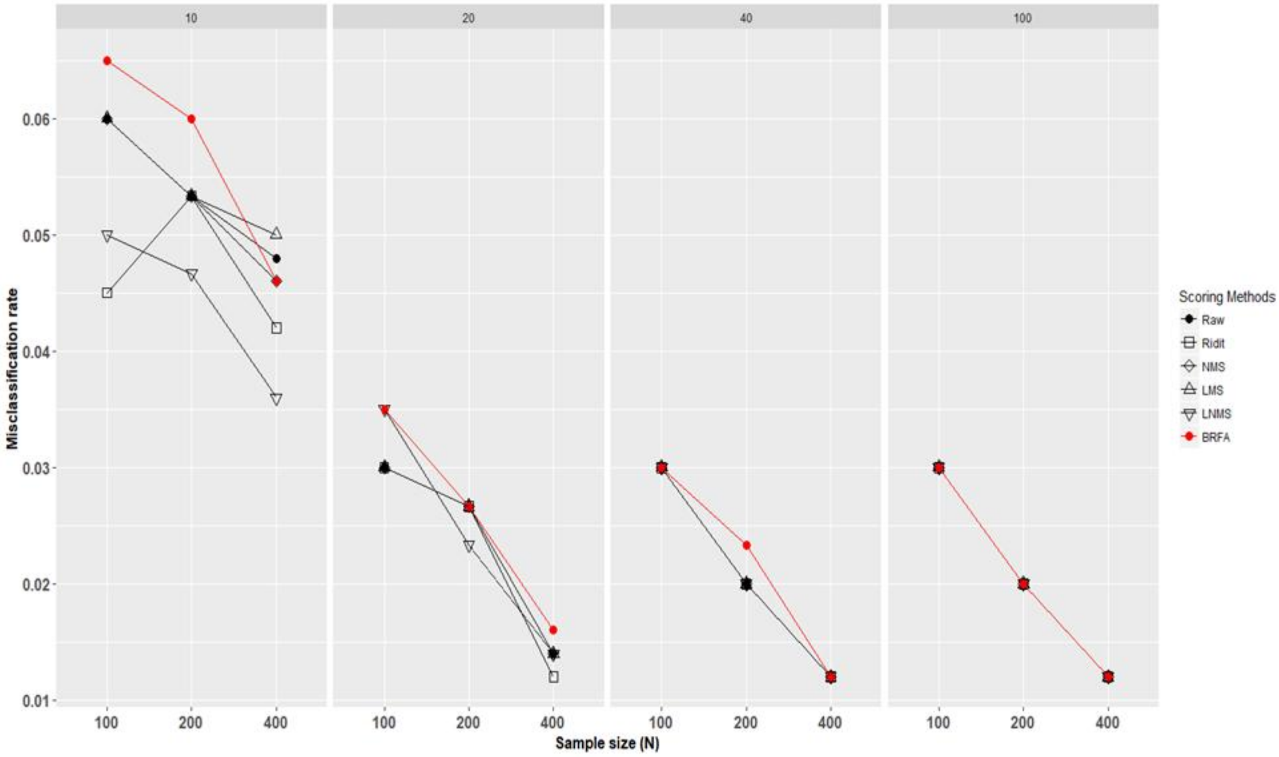

Figure A.47: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

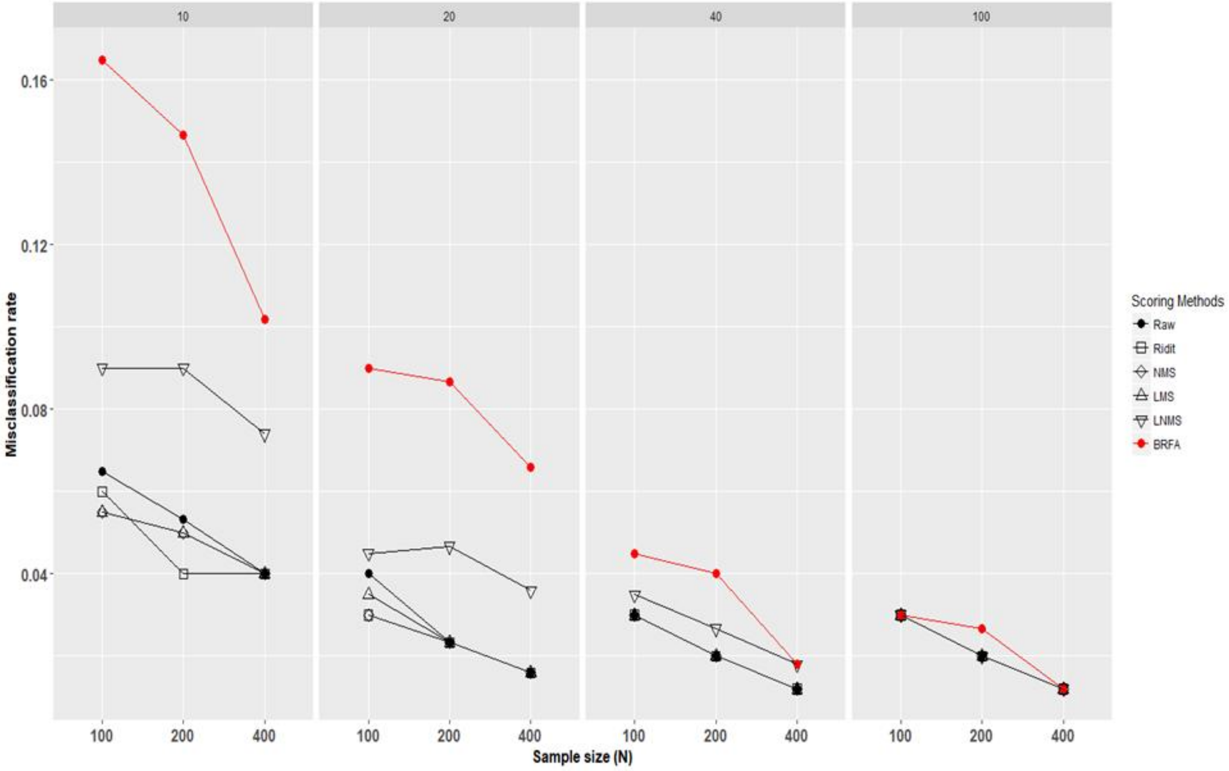

Figure A.48: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have $C = 3$ categories.*

Table A.13: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for naive Bayes in the four considered scenarios. Simulation results refer to the case of ordinal features with three categories ($C = 3$).*

	Scenario 1											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.20 (0.03)	0.22 (0.03)	0.22 (0.02)	0.22 (0.02)	0.15 (0.02)	0.10 (0.01)	0.06 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
Ridit	0.19 (0.04)	0.22 (0.03)	0.23 (0.02)	0.22 (0.02)	0.17 (0.02)	0.10 (0.01)	0.06 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
NM	0.20 (0.03)	0.22 (0.03)	0.21 (0.02)	0.25 (0.02)	0.15 (0.02)	0.10 (0.01)	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
Blom	0.20 (0.03)	0.22 (0.03)	0.21 (0.02)	0.25 (0.02)	0.15 (0.02)	0.10 (0.01)	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
NMS	0.20 (0.03)	0.22 (0.03)	0.22 (0.02)	0.23 (0.02)	0.15 (0.02)	0.10 (0.01)	0.06 (0.02)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)
LMS	0.20 (0.03)	0.22 (0.03)	0.21 (0.02)	0.24 (0.02)	0.15 (0.02)	0.10 (0.01)	0.07 (0.02)	0.06 (0.01)	0.04 (0.01)	0.05 (0.01)	0.03 (0.01)	0.04 (0.01)
LNMS	0.28 (0.05)	0.34 (0.02)	0.32 (0.01)	0.33 (0.03)	0.27 (0.03)	0.27 (0.02)	0.24 (0.02)	0.22 (0.02)	0.22 (0.01)	0.16 (0.03)	0.19 (0.02)	0.20 (0.01)
BRFA	0.26 (0.04)	0.30 (0.02)	0.27 (0.02)	0.25 (0.03)	0.23 (0.02)	0.12 (0.01)	0.09 (0.03)	0.08 (0.02)	0.07 (0.01)	0.04 (0.02)	0.04 (0.01)	0.03 (0.01)
	Scenario 2											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.07 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
NM	0.07 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
Blom	0.07 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
LMS	0.07 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
LNMS	0.21 (0.02)	0.19 (0.03)	0.23 (0.02)	0.12 (0.03)	0.13 (0.02)	0.14 (0.02)	0.06 (0.02)	0.06 (0.01)	0.05 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
BRFA	0.14 (0.03)	0.11 (0.02)	0.08 (0.01)	0.11 (0.02)	0.10 (0.02)	0.06 (0.01)	0.10 (0.02)	0.07 (0.01)	0.06 (0.01)	0.04 (0.01)	0.04 (0.01)	0.02 (0.01)
	Scenario 3											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
Ridit	0.04 (0.02)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.05 (0.01)	0.05 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
LNMS	0.05 (0.02)	0.05 (0.01)	0.04 (0.01)	0.04 (0.02)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
BRFA	0.06 (0.02)	0.06 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)
	Mixture scenario											
	p=10			**p=20**			**p=40**			**p=100**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.05 (0.02)	0.04 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
Ridit	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
NM	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
Blom	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
LMS	0.06 (0.02)	0.05 (0.02)	0.04 (0.01)	0.04 (0.02)	0.02 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
LNMS	0.09 (0.02)	0.09 (0.02)	0.07 (0.01)	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.04 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
BRFA	0.16 (0.03)	0.15 (0.02)	0.10 (0.01)	0.09 (0.02)	0.09 (0.02)	0.07 (0.01)	0.04 (0.02)	0.04 (0.01)	0.02 (0.01)	0.03 (0.01)	0.03 (0.01)	0.01 (0.01)

A.5.3 Simulation results for $C = 357$

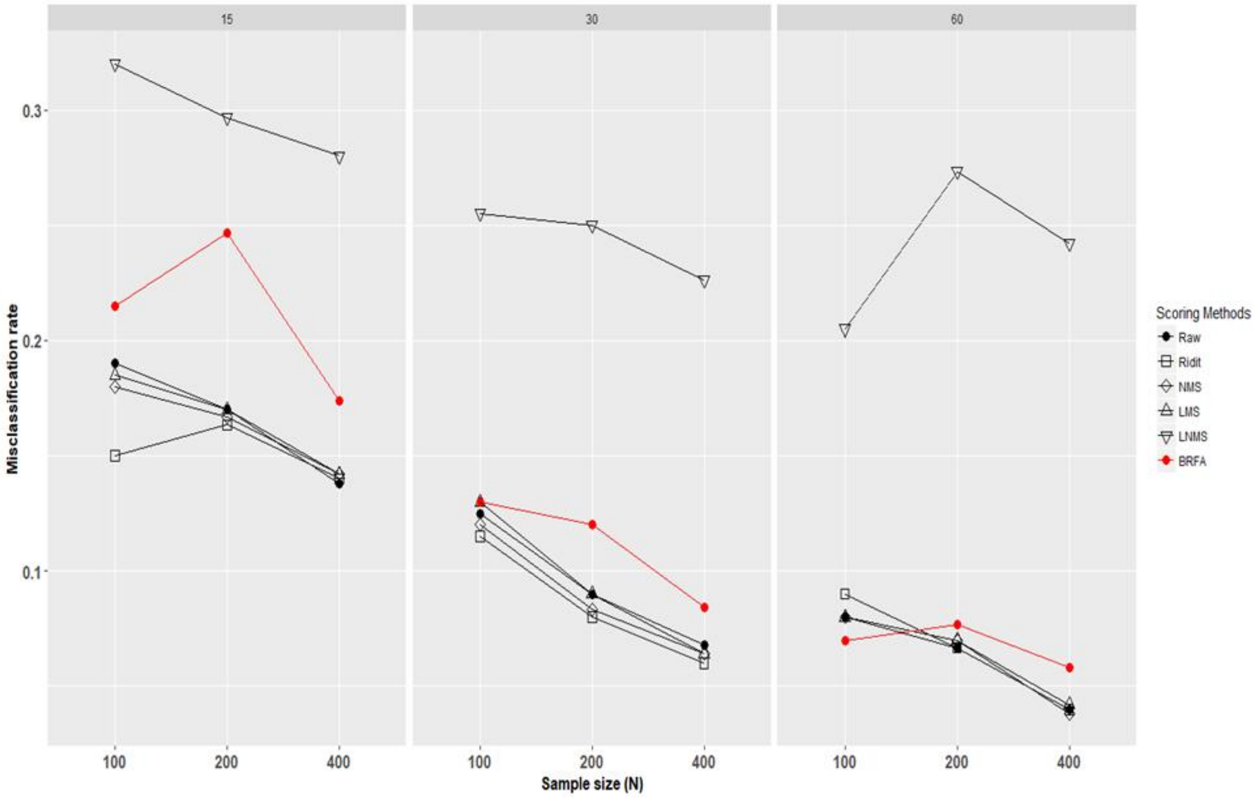

Figure A.49: *Results for scenario 1. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$*.*

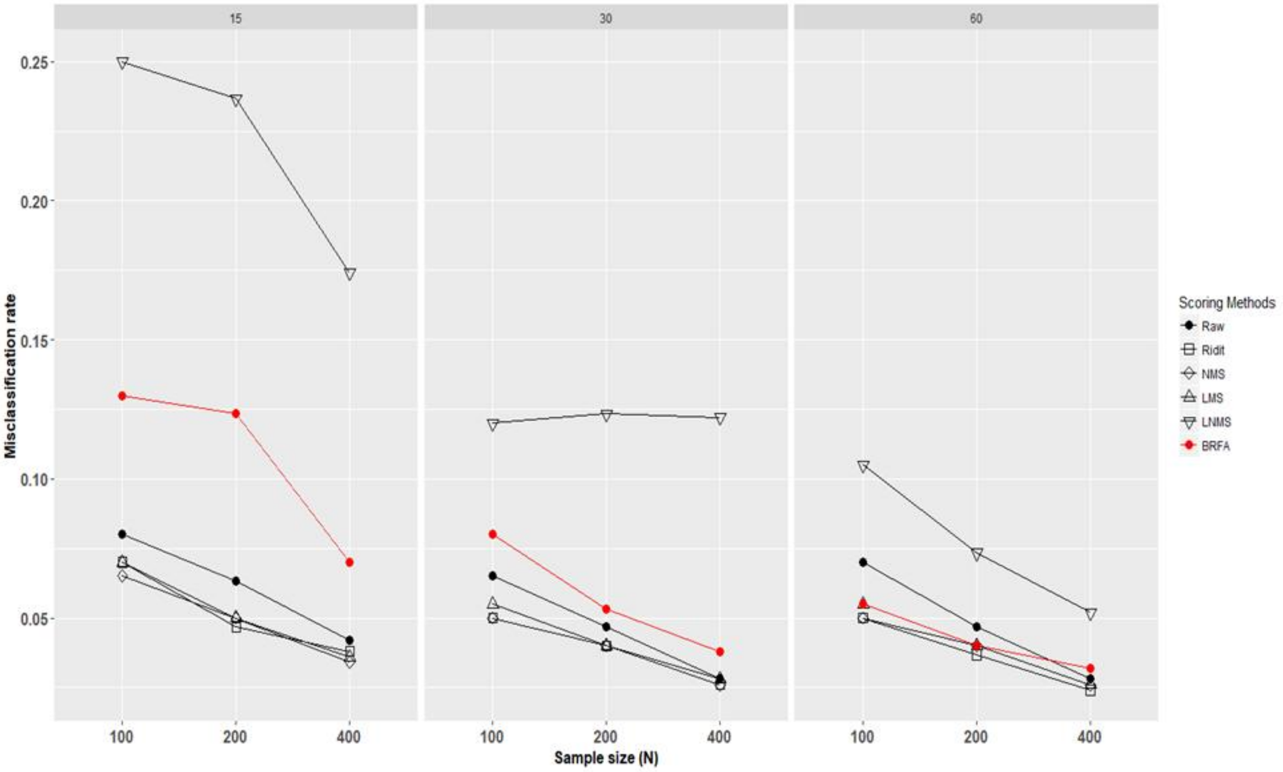

Figure A.50: *Results for scenario 2. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

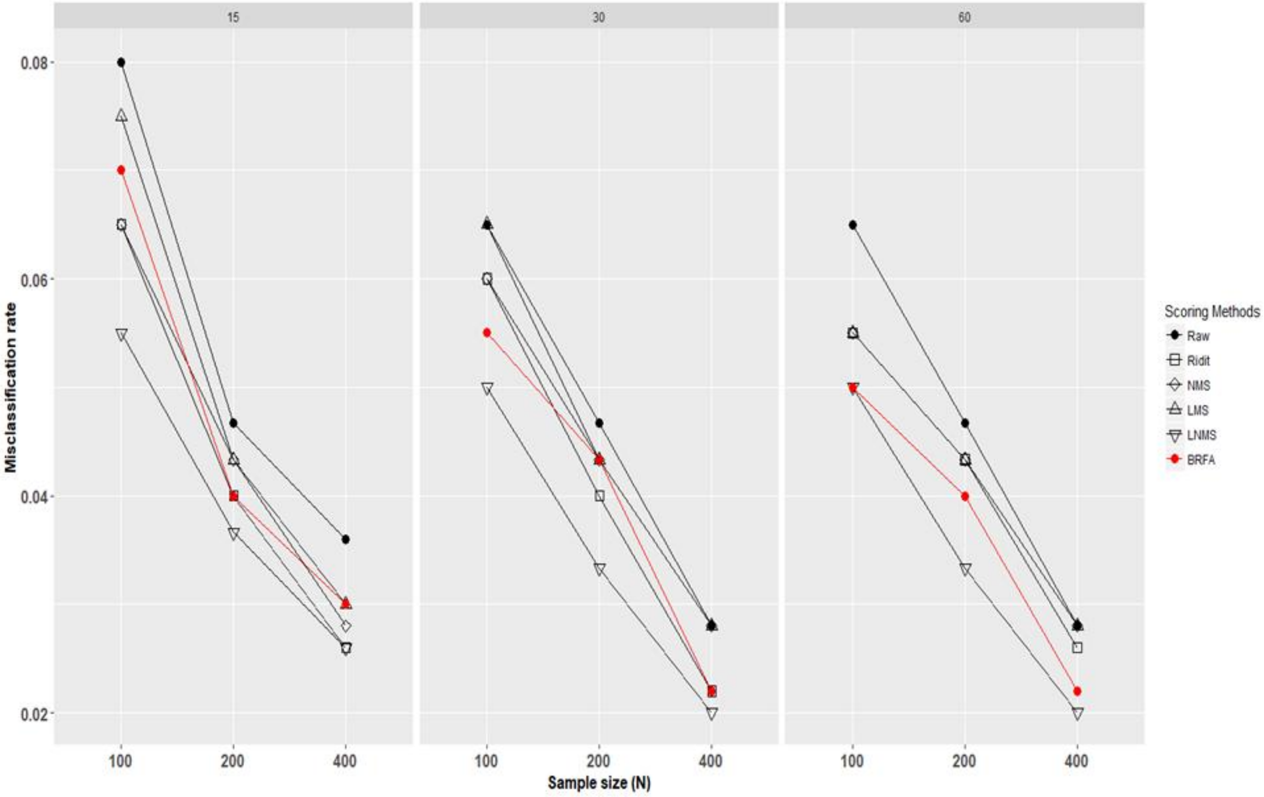

Figure A.51: *Results for scenario 3. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

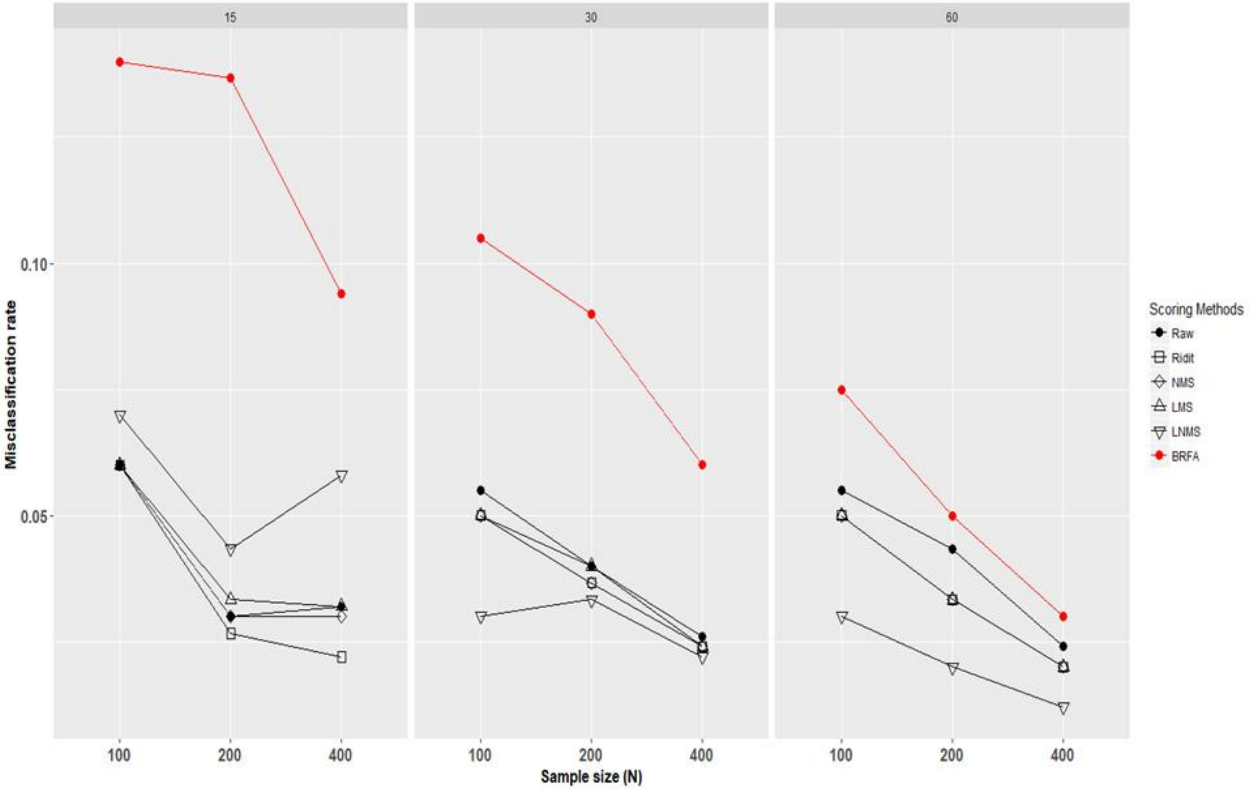

Figure A.52: *Results for "mixture" scenario. Each block corresponds to a different number of features. The points inside each block are the mean misclassification rates from a 10-fold cross-validation for different number of instances. Ordinal variables have been generated so that they have different number of categories* $C = 357$.

Table A.14: *Mean misclassification rates (with standard errors in brackets) over a 10-fold cross-validation for naive Bayes in the four considered scenarios. Simulation results refer to the case of ordinal features with different number of categories ($C = 357$).*

				Scenario 1					
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.19 (0.03)	0.17 (0.02)	0.14 (0.02)	0.12 (0.02)	0.09 (0.02)	0.07 (0.01)	0.08 (0.02)	0.07 (0.01)	0.04 (0.01)
Ridit	0.15 (0.02)	0.16 (0.02)	0.14 (0.02)	0.12 (0.02)	0.08 (0.02)	0.06 (0.01)	0.09 (0.02)	0.07 (0.01)	0.04 (0.01)
NM	0.16 (0.02)	0.16 (0.02)	0.15 (0.02)	0.12 (0.02)	0.08 (0.02)	0.06 (0.01)	0.08 (0.02)	0.07 (0.02)	0.04 (0.01)
Blom	0.16 (0.02)	0.16 (0.02)	0.15 (0.02)	0.12 (0.02)	0.08 (0.02)	0.06 (0.01)	0.08 (0.02)	0.07 (0.02)	0.04 (0.01)
NMS	0.18 (0.02)	0.17 (0.02)	0.14 (0.02)	0.12 (0.02)	0.08 (0.02)	0.06 (0.02)	0.08 (0.02)	0.07 (0.02)	0.04 (0.01)
LMS	0.18 (0.02)	0.17 (0.02)	0.14 (0.01)	0.13 (0.03)	0.09 (0.02)	0.06 (0.01)	0.08 (0.02)	0.07 (0.02)	0.04 (0.01)
LNMS	0.32 (0.01)	0.30 (0.02)	0.28 (0.02)	0.26 (0.03)	0.25 (0.03)	0.23 (0.02)	0.20 (0.02)	0.27 (0.03)	0.24 (0.01)
BRFA	0.22 (0.03)	0.25 (0.03)	0.17 (0.03)	0.13 (0.02)	0.12 (0.02)	0.08 (0.01)	0.07 (0.02)	0.08 (0.02)	0.06 (0.01)
				Scenario 2					
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.06 (0.02)	0.04 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.07 (0.02)	0.05 (0.01)	0.03 (0.01)
Ridit	0.07 (0.02)	0.05 (0.01)	0.04 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)
NM	0.07 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)
Blom	0.07 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)
NMS	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)	0.05 (0.01)	0.04 (0.01)	0.03 (0.01)
LMS	0.07 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)
LNMS	0.25 (0.03)	0.24 (0.03)	0.17 (0.03)	0.12 (0.02)	0.12 (0.02)	0.12 (0.01)	0.1 (0.02)	0.07 (0.01)	0.05 (0.01)
BRFA	0.13 (0.02)	0.12 (0.02)	0.07 (0.02)	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
				Scenario 3					
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.08 (0.02)	0.05 (0.01)	0.04 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)	0.06 (0.02)	0.05 (0.01)	0.03 (0.01)
Ridit	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
NM	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
Blom	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
NMS	0.06 (0.01)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
LMS	0.08 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)
LNMS	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
BRFA	0.07 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.01)	0.04 (0.01)	0.02 (0.01)	0.05 (0.01)	0.04 (0.01)	0.02 (0.01)
				Mixture scenario					
	p=15			**p=30**			**p=60**		
	N=100	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**	**N=100**	**N=200**	**N=400**
Raw	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.03 (0.01)	0.06 (0.02)	0.04 (0.01)	0.02 (0.01)
Ridit	0.06 (0.02)	0.03 (0.01)	0.02 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
NM	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
Blom	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
NMS	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LMS	0.06 (0.02)	0.03 (0.01)	0.03 (0.01)	0.05 (0.02)	0.04 (0.01)	0.02 (0.01)	0.05 (0.02)	0.03 (0.01)	0.02 (0.01)
LNMS	0.07 (0.02)	0.04 (0.01)	0.06 (0.01)	0.03 (0.01)	0.03 (0.01)	0.02 (0.01)	0.03 (0.01)	0.02 (0.01)	0.01 (0.01)
BRFA	0.14 (0.02)	0.14 (0.02)	0.09 (0.02)	0.1 (0.03)	0.09 (0.01)	0.06 (0.01)	0.08 (0.02)	0.05 (0.01)	0.03 (0.01)

Bibliography

Agresti, Alan. 2003. *Categorical data analysis*. Vol. 482. John Wiley & Sons.

Alcalá-Fdez, Jesús, Sanchez, Luciano, Garcia, Salvador, del Jesus, Maria Jose, Ventura, Sebastian, Garrell, Josep Maria, Otero, Jose, Romero, Cristóbal, Bacardit, Jaume, Rivas, Victor M, *et al.* 2009. KEEL: a software tool to assess evolutionary algorithms for data mining problems. *Soft Computing-A Fusion of Foundations, Methodologies and Applications*, **13**(3), 307–318.

Alcalá-Fdez, Jesús, Fernández, Alberto, Luengo, Julián, Derrac, Joaquín, García, Salvador, Sánchez, Luciano, & Herrera, Francisco. 2011. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, **17**.

Arminger, Gerhard, & Küsters, Ulrich. 1988. Latent trait models with indicators of mixed measurement level. *Pages 51–73 of: Latent trait and latent class models*. Springer.

Askey, Richard. 2005. The 1839 paper on permutations: its relation to the Rodrigues formula and further developments. *Mathematics and Social Utopias in France: Olinde Rodrigues and His Times*, **28**, 105–118.

Babolian, Esmail, MasjedJamei, Mohammad, & Eslahchi, MR. 2005. On numerical improvement of Gauss–Legendre quadrature rules. *Applied Mathematics and Computation*, **160**(3), 779–789.

Bartholomew, David J. 1980. Factor analysis for categorical data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 293–321.

Bartholomew, David J. 1981. Posterior analysis of the factor model. *British Journal of Mathematical and Statistical Psychology*, **34**(1), 93–99.

Bartholomew, David J. 1983. Latent variable models for ordered categorical data. *Journal of Econometrics*, **22**(1-2), 229–243.

Bartholomew, David J. 1984. Scaling binary data using a factor model. *Journal of the Royal Statistical Society. Series B (Methodological)*, 120–123.

Bartholomew, David J. 1988. The sensitivity of latent trait analysis to choice of prior distribution. *British Journal of Mathematical and Statistical Psychology*, **41**(1), 101–107.

Beauducel, Andre, & Herzberg, Philipp Yorck. 2006. On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, **13**(2), 186–203.

Biernacki, Christophe, & Jacques, Julien. 2016. Model-based clustering of multivariate ordinal data relying on a stochastic binary search algorithm. *Statistics and Computing*, **26**(5), 929–943.

Bishop, Christopher M. 2006. *Pattern recognition and machine learning*. springer.

Blom, Gunnar. 1958. *Statistical estimates and transformed beta-variables*. Ph.D. thesis, Almqvist & Wiksell.

Bock, RD, & Aitkin, M. 1982. Marginal maximum likelihood estimation of item parameters. *Psychometrika*, **47**(3), 369–369.

Bollen, K. A. 1989. *Introduction, in Structural Equations with Latent Variables*. John Wiley & Sons.

Brent, Richard P. 2013. *Algorithms for minimization without derivatives*. Courier Corporation.

Brockett, Patrick L. 1981. A note on the numerical assignment of scores to ranked categorical data.

Brockett, Patrick L, & Levine, Arnold. 1977. On a characterization of ridits. *The Annals of Statistics*, 1245–1248.

Bross, Irwin DJ. 1958. How to use ridit analysis. *Biometrics*, 18–38.

Cagnone, Silvia, & Viroli, Cinzia. 2012. A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, **12**(3), 257–277.

Cagnone, Silvia, & Viroli, Cinzia. 2014. A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. *AStA Advances in Statistical Analysis*, **98**(1), 1–20.

Chen, HC, & Wang, NS. 2014. The assignment of scores procedure for ordinal categorical data. *TheScientificWorldJournal*, **2014**, 304213–304213.

Clogg, Clifford C. 1982. Some models for the analysis of association in multiway cross-classifications having ordered categories. *Journal of the American Statistical Association*, **77**(380), 803–815.

Cortes, Corinna, & Vapnik, Vladimir. 1995. Support-vector networks. *Machine learning*, **20**(3), 273–297.

Dabney, Alan R. 2005. Classification of microarrays to nearest centroids. *Bioinformatics*, **21**(22), 4148–4154.

Dempster, Arthur P, Laird, Nan M, & Rubin, Donald B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Dennis, JE, & Woods, Daniel J. 1987. Optimization on microcomputers: The Nelder-Mead simplex algorithm. *New computing environments: microcomputers in large-scale computing*, **11**, 6–122.

Fan, Jianqing, & Fan, Yingying. 2008. High dimensional classification using features annealed independence rules. *Annals of statistics*, **36**(6), 2605.

Fielding, Antony. 1993. Scoring functions for ordered classifications in statistical analysis. *Quality & Quantity*, **27**(1), 1–17.

Fielding, Antony. 1997. On scoring ordered classifications. *British Journal of Mathematical and Statistical Psychology*, **50**(2), 285–307.

Fielding, Antony. 1999. Why use arbitrary points scores?: ordered categories in models of educational progress. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **162**(3), 303–328.

Fisher, Ronald A. 1936. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, **7**(2), 179–188.

Fisher, Ronald A. 1940. The precision of discriminant functions. *Annals of Human Genetics*, **10**(1), 422–429.

Friedman, Jerome, Hastie, Trevor, & Tibshirani, Robert. 2001. *The elements of statistical learning.* Vol. 1. Springer series in statistics New York.

Galton, Francis. 1902. The most suitable proportion between the value of first and second prizes. *Biometrika*, 385–399.

Goldstein, Harvey. 2011. *Multilevel statistical models.* Vol. 922. John Wiley & Sons.

Greene, William H. 1993. Econometric Analysis (2th edition). *Macmillan, NJ.*

Hall, Peter, Titterington, DM, & Xue, Jing-Hao. 2009. Median-based classifiers for high-dimensional data. *Journal of the American Statistical Association*, **104**(488), 1597–1608.

Hand, David J. 1997. Construction and Assessment of Classification Rules. *Statistics in medicine*, **20**, 326–327.

Harter, H Leon. 1961. Expected values of normal order statistics. *Biometrika*, **48**(1/2), 151–165.

Hastie, Trevor, Tibshirani, Robert, & Friedman, Jerome. 2009. Overview of supervised learning. *Pages 9–41 of: The elements of statistical learning.* Springer.

Hennig, Christian, & Viroli, Cinzia. 2016a. Quantile-based classifiers. *Biometrika*, **103**(2), 435–446.

Hennig, Christian, & Viroli, Cinzia. 2016b. *quantileDA: Quantile Classifier.* R package version 1.1.

Hollander, Myles, Wolfe, Douglas A, & Chicken, Eric. 2013. *Nonparametric statistical methods.* John Wiley & Sons.

Hyndman, Rob J, & Fan, Yanan. 1996. Sample quantiles in statistical packages. *The American Statistician*, **50**(4), 361–365.

Iannario, Maria. 2010. On the identifiability of a mixture model for ordinal data. *Metron*, **68**(1), 87–94.

Iannario, Maria, & Piccolo, Domenico. 2012. CUB models: Statistical methods and empirical evidence. *Modern Analysis of Customer Surveys: with applications using R*, 231–258.

Irincheeva, Irina, Cantoni, Eva, & Genton, Marc G. 2012. Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, **39**(4), 663–680.

Ivory, James. 1824. On the figure requisite to maintain the equilibrium of a homogeneous fluid mass that revolves upon an axis. *Philosophical Transactions of the Royal Society of London*, **114**, 85–150.

Jacobi, JDG. 1827. Ueber eine besondere Gattung algebraischer Functionen, die aus der Entwicklung der Function (1-2xz+ z2)... entstehen. *Journal für die reine und angewandte Mathematik*, **2**, 223–226.

Jacques, Julien, & Biernacki, Christophe. 2017. Model-based co-clustering for ordinal data.

Jarratt, P. 1967. An iterative method for locating turning points. *The Computer Journal*, **10**(1), 82–84.

John, George H, & Langley, Pat. 1995. Estimating continuous distributions in Bayesian classifiers. *Pages 338–345 of: Proceedings of the Eleventh conference on Uncertainty in artificial intelligence.* Morgan Kaufmann Publishers Inc.

Johnson, David Richard, & Creech, James C. 1983. Ordinal measures in multiple indicator models: A simulation study of categorization error. *American Sociological Review*, 398–407.

Jöreskog, Karl G. 1990. New developments in LISREL: Analysis of ordinal variables using polychoric correlations and weighted least squares. *Quality & Quantity*, **24**(4), 387–404.

Jöreskog, Karl G. 1994. On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika*, **59**(3), 381–389.

Jöreskog, Karl G, & Moustaki, Irini. 2001. Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research,*, **36**(3), 347–387.

Jörnsten, Rebecka. 2004. Clustering and classification based on the L1 data depth. *Journal of Multivariate Analysis*, **90**(1), 67–89.

Kampen, Jarl, & Swyngedouw, Marc. 2000. The ordinal controversy revisited. *Quality & Quantity*, **34**(1), 87–102.

Kiefer, Jack. 1953. Sequential minimax search for a maximum. *Proceedings of the American mathematical society*, **4**(3), 502–506.

Knott, Martin, & Bartholomew, David J. 1999. *Latent variable models and factor analysis*. Edward Arnold.

Labovitz, Sanford. 1970. The assignment of numbers to rank order categories. *American Sociological Review*, 515–524.

Labovitz, Sanford. 1971. In defense of assigning numbers to ranks. *American Sociological Review*, **36**(3), 521–522.

Lagarias, Jeffrey C, Reeds, James A, Wright, Margaret H, & Wright, Paul E. 1998. Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal on optimization*, **9**(1), 112–147.

Langley, Pat, Iba, Wayne, Thompson, Kevin, *et al.* 1992. An analysis of Bayesian classifiers. *Pages 223–228 of: Aaai*, vol. 90.

Lee, Sik-Yum, Poon, Wai-Yin, & Bentler, PM. 1990. Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & probability letters*, **9**(1), 91–97.

Lee, Sik-Yum, Poon, Wai-Yin, & Bentler, Peter M. 1995. A two-stage estimation of structural equation models with continuous and polytomous variables. *British Journal of Mathematical and Statistical Psychology*, **48**(2), 339–358.

Lichman, M. 2013. *UCI Machine Learning Repository*.

Lord, Frederic M, & Novick, Melvin R. 2008. *Statistical theories of mental test scores*. IAP.

Marcus-Roberts, Helen M, & Roberts, Fred S. 1987. Meaningless statistics. *Journal of Educational Statistics*, **12**(4), 383–394.

McCullagh, Peter. 1980. Regression models for ordinal data. *Journal of the royal statistical society. Series B (Methodological)*, 109–142.

McCullagh, Peter. 1984. Generalized linear models. *European Journal of Operational Research*, **16**(3), 285–292.

McLachlan, Geoffrey, & Krishnan, Thriyambakam. 2007. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons.

Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas, & Leisch, Friedrich. 2015. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-7.

Montanari, Angela, & Viroli, Cinzia. 2010. A skew-normal factor model for the analysis of student satisfaction towards university courses. *Journal of Applied Statistics*, **37**(3), 473–487.

Moustaki, Irini. 1996. A latent trait and a latent class model for mixed observed variables. *British journal of mathematical and statistical psychology*, **49**(2), 313–334.

Moustaki, Irini. 2000. A latent variable model for ordinal variables. *Applied psychological measurement*, **24**(3), 211–223.

Moustaki, Irini. 2003. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology*, **56**(2), 337–357.

Moustaki, Irini, & Knott, Martin. 2000. Generalized latent trait models. *Psychometrika*, **65**(3), 391–411.

Muthén, Bengt. 1984. A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49**(1), 115–132.

Nelder, John A, & Mead, Roger. 1965. A simplex method for function minimization. *The computer journal*, **7**(4), 308–313.

O'Brien, Robert M. 1979. The use of Pearson's with ordinal data. *American Sociological Review*, 851–857.

Pearson, Karl. 1902. Note on Francis Galton's problem. *Biometrika*, **1**(4), 390–99.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rhemtulla, Mijke, Brosseau-Liard, Patricia É, & Savalei, Victoria. 2012. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological methods*, **17**(3), 354.

Rodrigues, Olinde. 1815. *De l'attraction des sphéroïdes*. Ph.D. thesis.

Selvin, Steve. 1977. A further note on the interpretation of ridit analysis. *American Journal of Epidemiology*, **105**(1), 16–20.

Shi, Yuan, Li, Wenzhe, & Sha, Fei. 2016. Metric Learning for Ordinal Data. *Pages 2030–2036 of: AAAI*.

Snell, EJ. 1964. A scaling procedure for ordered categorical data. *Biometrics*, 592–607.

Solomon, Shira R, & Sawilowsky, Shlomo S. 2009. Impact of rank-based normalizing transformations on the accuracy of test scores.

Spearman, Charles. 1904. " General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, **15**(2), 201–292.

Stevens, Stanley Smith, *et al.* 1946. On the theory of scales of measurement.

Van Der Putten, Peter, & van Someren, Maarten. 2000. CoIL challenge 2000: The insurance company case. *Published by Sentient Machine Research, Amsterdam. Also a Leiden Institute of Advanced Computer Science Technical Report*, **9**, 1–43.

Van Der Putten, Peter, & Van Someren, Maarten. 2004. A bias-variance analysis of a real world learning problem: The CoIL challenge 2000. *Machine Learning*, **57**(1), 177–195.

Venables, W. N., & Ripley, B. D. 2002. *Modern Applied Statistics with S*. Fourth edn. New York: Springer. ISBN 0-387-95457-0.

Wall, Melanie M, Park, Jung Yeon, & Moustaki, Irini. 2015. IRT modeling in the presence of zero-inflation with application to psychiatric disorder severity. *Applied Psychological Measurement*, **39**(8), 583–597.

Wedel, Michel, & Kamakura, Wagner A. 2001. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, **66**(4), 515–530.

Williams, EJ. 1952. Use of scores for the analysis of association in contingency tables. *Biometrika*, **39**(3/4), 274–289.

Yates, Frank. 1948. The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, **35**(1/2), 176–181.

Yung, Yiu-Fai. 1997. Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, **62**(3), 297–330.