UNIVERSIDAD DE OVIEDO

MASTER IN SOFT COMPUTING
AND INTELLIGENT DATA ANALYSIS

MASTER PROJECT

INTELLIGENT DATA ANALYSIS ON A VISUAL PERCEPTION EXPERIMENT

Luis Pedro Hidalgo Martín
July 2012

Advisor

Prof. Gil González Rodríguez

This thesis was typeset using LaTeX. The main typeface is Palatino, which stems from the humanist fonts of the Italian Renaissance. Typographical decisions were based on *The Elements of Typographic Style* by Robert Bringhust. The overall design of the document, with a simple structure and extensive usage of sidenotes, was inspired by the work of Edward Tufte.

Final version, July 2012

Contents

I would like to express my deeply felt thanks to Prof. Gil González Rodríguez for all the hours he has dedicated, his insightful comments, thoughtful guidance, and ideas during the elaboration of this thesis.

Abstract

This thesis deals with supervised classification of fuzzy data obtained from a random experiment. The data generation process is modeled using three approaches. First, a naive direct multivariate method which treats the data *as is*. Second, a compositional data transformation that views data as points in a Simplex space. Finally, we also employ an algorithm that relies on random fuzzy sets.

The first two approaches have been tested on a classical setting of supervised classifiers. The fuzzy approach has been tested on a custom family of classifiers for fuzzy data. Two of the fuzzy algorithms are novel contributions.

The empirical test consists on two experiments. One concerning fuzzy perceptions and linguistic labels, and the other concerning fuzzy perceptions and the gender of the individual that generated the perceptions.

Introduction

Problem setting

This thesis presents a study in supervised classification, and builds upon the theories and models developed at the Statistical Methods with Imprecise Random Elements [1] research group, located at the University of Oviedo. More concretely, this work continues a recent effort by the thesis advisor's and close collaborators[2]. The dissertation is centered around a visual perception experiment, whose outcomes are modeled using fuzzy theory.

[1] SMIRE, http://bellman.ciencias.uniovi.es/SMIRE

[2] A. Colubi, G. González-Rodríguez, M. Ángeles Gil, and W. Trutschnig. Nonparametric criteria for supervised classification of fuzzy data. *International Journal of Approximate Reasoning*, 2011

Goals

To broaden a previous supervised classification problem against a set of classical classifiers. To analyze a new supervised classification problem with the same batch of classifiers. Perform the same analysis with a fully fuzzy space theory framework. And to improve some of the fuzzy classifiers methods provided in the literature.

Outline

Fuzzy Data Treated as Functional Data We provide a brief description of the theory for working with fuzzy data in a fully coherent way. With that a coherent arithmetic, inner product and distance could be used in a fully fuzzy setting and that provides the basic tools for doing inference in a fuzzy setting.

Classification of Fuzzy Convex Data A description of Colubi et al.[2] previous supervised classification algorithms is provided. We also present a contribution in the form of two new algorithms for the fuzzy setting.

A Visual Perception Experiment Next we describe the experiment upon this project builds its analysis.

Fuzzy Data Treated as Compositional Data We provide another way to analyze the data besides the direct multivariate—*as is*— manner and the fuzzy framework. Due to the nature of the data, which is constrained, a mapping between the data an a compositional datum could be established. This provides another set of tools for analyzing the data. A brief overview of the mathematical framework for dealing with compositional data is presented.

Classification Analysis We complement the original classification problem testing against a batch of classifiers and the previous algorithms developed explicitly for this kind of fuzzy data plus new ones developed for this work. We try to solve a new classification problems: try to infer the gender of the respondent. The analysis is done under three settings, (direct) Multivariate —data *as is*—, Simplex space —the data is mapped into a composition—, Fuzzy space —the data is treated in a coherent fuzzy setting—. A set of classical classifiers are used against the two first representations, and a family of fuzzy classifiers for the fuzzy setting.

1
Fuzzy Data Treated as Functional Data

The exposition presented here follows the one provided in González-Rodríguez et al. 2010, with additional diagrams and material from Ángeles Gil 2010.

1.1 Introduction and motivation

Data which cannot be exactly described by means of numerical values, such as evaluations, medical diagnosis or quality ratings, to name but a few, are frequently classified as either nominal or ordinal. A well-known example is the so-called Lickert[1] scales in which categories are labeled with numerical values. Using these scales, the statistical analysis is limited. Many parameters and techniques cannot be directly used or, when they can, the interpretation and reliability of the conclusions are considerably reduced. Additionally, the transition from one category to another is rather abrupt. A third concern is that categories are not perceived in the same manner by different observers, so that variability and accuracy cannot always be well captured.

A easy-to-use representation of such data through fuzzy values is to be considered. The measurement scale of fuzzy values includes, in particular, real vectors and set values as special elements. It is more expressive than ordinal scales and more accurate than rounding or using real or vectorial-valued codes. The arithmetic and metric to be used make it possible to extend naturally many of the usual statistical measures and techniques. The transition between closely different values can be made gradually, and the variability, accuracy and possible subjectiveness can be well reflected in describing data.

[1] R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932; and I.E. Allen and C.A. Seaman. Likert scales and data analyses. *Quality Progress*, 40 (7):64–65, 2007

1.2 Fuzzy data viewed as functional data

1.2.1 Fuzzy compact convex sets

Let $\mathcal{K}_c(\mathbb{R}^p)$ be the class of the non-empty compact convex subsets of $\mathbb{R}^p$ and let $\mathcal{F}_c(\mathbb{R}^p)$ denote the class of normal and convex upper semicontinuous fuzzy sets of $\mathbb{R}^p$ with bounded closure of the support, that is

$$\mathcal{F}_c(\mathbb{R}^p) = \{\widetilde{U} : \mathbb{R}^p \to [0,1] \mid \widetilde{U}_\alpha \in \mathcal{K}_c(\mathbb{R}^p) \quad \forall \alpha \in (0,1]\}$$

with

$$\widetilde{U}_\alpha = \{x \in \mathbb{R}^p | \widetilde{U}(x) \geq \alpha, \ \forall \alpha \in (0,1]\}$$

$$\widetilde{U}_0 = cl(\{x \in \mathbb{R}^p | \widetilde{U}(x) > 0\})$$

These class of fuzzy sets are also called *fuzzy values*, when $p = 1$ the fuzzy values are referred to as fuzzy numbers.

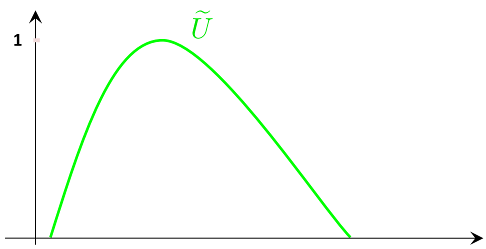

Figure 1.1: Fuzzy value in $\mathcal{F}_c(\mathbb{R})$

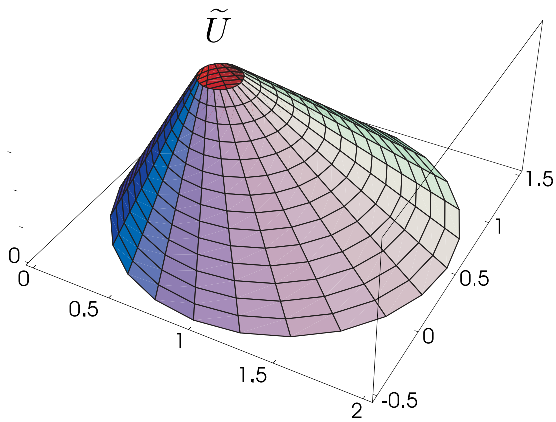

Figure 1.2: Fuzzy value in $\mathcal{F}_c(\mathbb{R}^2)$

Formally, fuzzy values are $[0,1]$-valued upper semicontinuous functions with non-empty convex bounded α-levels. Real, vectorial, interval and set-valued data can be viewed as particular fuzzy data, by identifying them with the associated indicator functions.

1.2.2 *Arithmetic operations*

The space $\mathcal{F}_c(\mathbb{R}^p)$ can be endowed with an inner composition law extending the Minkowski [2] addition between sets and an external one which is the product by a scalar. These laws are compatible with the ones obtained by applying Zadeh's extension principle[3].

[2] $A + B = \{a + b : a \in A, b \in B\}$
$\lambda A = \{\lambda a : a \in A\}$ $\lambda \in \mathbb{R}$

[3] L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning–I. *Information sciences*, 8(3):199–249, 1975

Given $\widetilde{U}_\alpha, \widetilde{V}_\alpha \in \mathcal{F}_c(\mathbb{R}^p)$, and $\gamma \in \mathbb{R}$,

$$(\widetilde{U} + \widetilde{V})_\alpha = \widetilde{U}_\alpha + \widetilde{V}_\alpha = \{y + z : y \in \widetilde{U}_\alpha, z \in \widetilde{V}_\alpha\}$$
$$= \left[\inf \widetilde{U}_\alpha + \inf \widetilde{V}_\alpha, \sup \widetilde{U}_\alpha + \sup \widetilde{V}_\alpha\right]$$

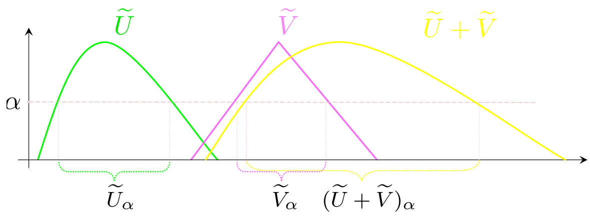

Figure 1.3: Fuzzy sum in $\mathcal{F}_c(\mathbb{R})$

$$(\gamma\widetilde{U})_\alpha = \gamma\widetilde{U}_\alpha = \{\gamma y : y \in \widetilde{U}_\alpha\} = \begin{cases} \left[\gamma \cdot \inf \widetilde{U}_\alpha, \gamma \cdot \sup \widetilde{U}_\alpha\right] & \text{if } \gamma \geq 0 \\ \left[\gamma \cdot \sup \widetilde{U}_\alpha, \gamma \cdot \inf \widetilde{U}_\alpha\right] & \text{if } \gamma < 0 \end{cases}$$

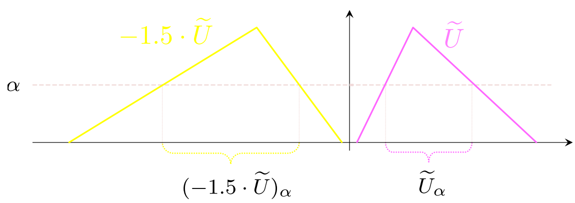

Figure 1.4: Fuzzy multiplication by an scalar in $\mathcal{F}_c(\mathbb{R})$

This arithmetic does not coincide with the usual one for functions. The application of the functional arithmetic in $\mathcal{F}_c(\mathbb{R}^p)$ may lead to elements out of this space, and the fuzzy meaning would be lost. Thus, from a formal point of view fuzzy data are a special kind of functional data. Nevertheless, although the considered arithmetic is quite natural and intuitive to interpret in the setting of fuzzy sets it does not coincide with the usual one for functional data. the space $(\mathcal{F}_c(\mathbb{R}^p), +, \cdot)$ has not a linear (but a semilinear-conical) structure, because the sum extends level-wise the Minkowski sum of sets.

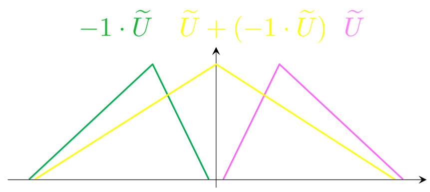

Figure 1.5: Semilinear structure of $(\mathcal{F}_c(\mathbb{R}), +, \cdot)$ $\widetilde{U} + (-1 \cdot \widetilde{U}) \neq 1_{\{0\}}$

A fuzzy value $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ models an ill-defined subset of $\mathbb{R}^p$, so that for each $x \in \mathbb{R}^p$ the value $\widetilde{U}(x)$ can be interpreted as 'degree of membership' of x to $\widetilde{U}$. Alternatively, $\widetilde{U}$ may be interpreted as the 'degree of compatibility' of x with an ill-defined property $\widetilde{U}$. In practice, fuzzy data usually come from either a pre-established classification, such as the danger of forest fires[4], or from a designed experiment. This is the case of the expert evaluation of the trees in a reforestation analyzed by Colubi[5], where the ill-defined characteristic 'quality' is individually described through a fuzzy set. Obviously, accuracy and variability of data are much better captured by using individual fuzzy assessments than by considering a pre-fixed list of fuzzy values.

[4] A. Colubi and G. González-Rodríguez. Triangular fuzzification of random variables and power of distribution tests: Empirical discussion. *Computational statistics & data analysis*, 51 (9):4742–4750, 2007

[5] A. Colubi. Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy-and real-valued data. *Fuzzy Sets and Systems*, 160(3):344–356, 2009

1.2.3 *The support function: a functional representation of fuzzy values*

Consider the space $\mathcal{H} = \mathcal{L}^2(\mathbb{S}^{p-1} \times (0,1], \lambda_p \times \lambda)$ of the L^2-type real-valued functions defined on the unit sphere $\mathbb{S}^{p-1}$ of $\mathbb{R}^p$ times the interval $(0,1]$ with respect to the corresponding normalized Lebesgue measures denoted by λ_p and λ. The mid/spr *decomposition of a function* $f \in \mathcal{H}$ can be defined [6] as

$f = \operatorname{mid} f + \operatorname{spr} f$ where, for all $u \in \mathbb{S}^{p-1}$ and $\alpha \in (0,1]$,

$$\operatorname{mid} f(u,\alpha) = \frac{f(u,\alpha) - f(-u,\alpha)}{2}, \quad \operatorname{spr} f(u,\alpha) = \frac{f(u,\alpha) + f(-u,\alpha)}{2},$$

where if $u = (u_1, \ldots, u_p) \in \mathbb{S}^{p-1}$, $-u$ denotes the element

[6] G. González-Rodríguez, A. Colubi, and M.Á. Gil. Fuzzy data treated as functional data: a one-way anova test approach. *Computational Statistics & Data Analysis*, 2010

$(-u_1, \ldots, -u_p) \in \mathbb{S}^{p-1}$. It can be proven that $\operatorname{mid} f,\ \operatorname{spr} f \in \mathcal{H}$, $\operatorname{mid} f$ is an odd function and $\operatorname{spr} f$ is an even one, w.r.t. the first component.

On this basis, a valuable inner product in $\mathcal{H}$ can be defined. More precisely, let $\theta \in (0, +\infty)$ [7] and let φ be a weighting measure formalized as an absolutely continuous probability measure on $([0,1], \mathcal{B}_{[0,1]})$ with positive mass function in $(0,1)$. For $f, g \in \mathcal{H}$ consider the value

[7] cf. 2.2 for an alternative definition

$$\langle f, g \rangle_\theta^\varphi = \left[\operatorname{mid} f, \operatorname{mid} g\right]^\varphi + \theta\left[\operatorname{spr} f, \operatorname{spr} g\right]^\varphi,$$

where

$$[f, g]^\varphi = \int_{(0,1]} \int_{\mathbb{S}^{p-1}} f(u,\alpha) g(u,\alpha)\, \mathrm{d}\lambda_p(u)\, \mathrm{d}\varphi(\alpha).$$

Then, the following properties are satisfied:

i) $\langle f, g \rangle_\theta^\varphi$ is an inner product in $\mathcal{H}$, for which the associated norm is denoted by $\| \cdot \|_\theta^\varphi$.

ii) The mid/spr decomposition of a function $f \in \mathcal{H}$ is orthogonal.

iii) $(\mathcal{H}, \langle \cdot, \cdot \rangle_\theta^\varphi)$ is a separable Hilbert space.

The *support function*[8] of $\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p)$ extends level-wise the notion of the support function of a set[9] and it is given by the mapping $s_{\widetilde{U}} : \mathbb{S}^{p-1} \times (0,1] \to \mathbb{R}$ defined so that

[8] M.L. Puri and D.A. Ralescu. The concept of normality for fuzzy random variables. *The Annals of Probability*, 13(4):1373–1379, 1985

$$s_{\widetilde{U}}(u,\alpha) = \sup_{v \in \widetilde{U}_\alpha} \langle u, v \rangle$$

[9] C. Castaing, M. Valadier, and SpringerLink (Service en ligne). *Convex analysis and measurable multifunctions*, volume 580. Springer-Verlag Berlin, 1977

for all $u \in \mathbb{S}^{p-1}, \alpha \in (0,1]$, where $\langle \cdot, \cdot \rangle$ denotes the inner product on $\mathbb{R}^p$. In general, one can state that $s_{\widetilde{U}}(u,\alpha)$ represents the "oriented" distance from $\mathbf{0} \in \mathbb{R}^p$ to the supporting hyperplane of $\widetilde{U}_\alpha$ which is orthogonal to u.

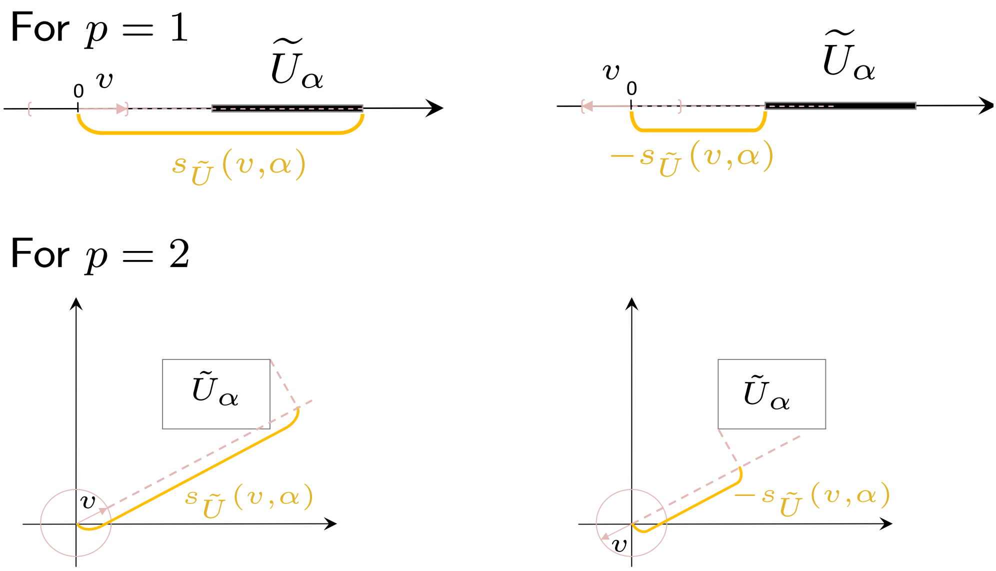

Figure 1.6: Support functions in $\mathcal{F}_c(\mathbb{R}^p)$

The mapping $s : \mathcal{F}_c^2(\mathbb{R}^p) \to \mathcal{H}$, such that $s(\widetilde{U}) = s_{\widetilde{U}}$ for all $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p) = \{\widetilde{U} \in \mathcal{F}_c(\mathbb{R}^p) : s_{\widetilde{U}} \in \mathcal{H}\}$, is semilinear, that is, it transform the fuzzy arithmetic to the functional arithmetic in the corresponding cone.

Let $\widetilde{U} \in \mathcal{F}_c^2(\mathbb{R}^p)$, then, from Trutschnig et al[10]. we have that

i) for all $\alpha \in (0,1]$ the projection of $\widetilde{U}_\alpha$ over a direction $u \in \mathbb{S}^{p-1}$ is given by the interval

$$\Pi_u \widetilde{U}_\alpha = \left[-s_{\widetilde{U}}(-u,\alpha), s_{\widetilde{U}}(u,\alpha)\right];$$

ii) $\operatorname{mid} s_{\widetilde{U}}(u,\alpha) = \dfrac{s_{\widetilde{U}}(u,\alpha) - s_{\widetilde{U}}(-u,\alpha)}{2}$ = mid-point/center of $\Pi_u \widetilde{U}_\alpha$;

iii) $\operatorname{spr} s_{\widetilde{U}}(u,\alpha) = \dfrac{s_{\widetilde{U}}(u,\alpha) + s_{\widetilde{U}}(-u,\alpha)}{2}$ = spread/radius of $\Pi_u \widetilde{U}_\alpha$;

iv) if $\operatorname{mid} \widetilde{U}(\bullet,\star)$ = mid-point of $\Pi_\bullet \widetilde{U}_\star$, $\operatorname{spr} \widetilde{U}(\bullet,\star)$ = spread of $\Pi_\bullet \widetilde{U}_\star$, then

$$s_{\widetilde{U}} = \operatorname{mid} \widetilde{U} + \operatorname{spr} \widetilde{U}.$$

[10] W. Trutschnig, G. González-Rodríguez, A. Colubi, and M.Á. Gil. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences*, 179(23): 3964–3972, 2009

Consequently, the mid and the spr of a fuzzy value can be interpreted as a kind of functional measurements of its 'location' and 'shape', respectively.

1.2.4 *Distance between fuzzy values and isometrical embedding*

The above-established connection induces a family of L^2 metrics on $\mathcal{F}_c^2(\mathbb{R}^p)$ from that associated with the norms $\|\cdot\|_\theta^\varphi$ on $\mathcal{H}$. Specifically, Trutschnig et al[11]. have introduced the following family of metrics.

[11] W. Trutschnig, G. González-Rodríguez, A. Colubi, and M.Á. Gil. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences*, 179(23): 3964–3972, 2009

Let $\theta \in (0, +\infty)$ and let φ be an absolutely continuous probability measure on $([0,1], \mathcal{B}_{[0,1]})$ with positive mass function in $(0,1)$. Then, the mapping $D_\theta^\varphi : \mathcal{F}_c^2(\mathbb{R}^p) \times \mathcal{F}_c^2(\mathbb{R}^p) \to [0, +\infty)$ such that for any $\widetilde{U}, \widetilde{V} \in \mathcal{F}_c^2(\mathbb{R}^p)$

$$\left(D_\theta^\varphi(\widetilde{U}, \widetilde{V})\right)^2 = \langle s_{\widetilde{U}} - s_{\widetilde{V}}, s_{\widetilde{U}} - s_{\widetilde{V}} \rangle_\theta^\varphi = \left(\|s_{\widetilde{U}} - s_{\widetilde{V}}\|_\theta^\varphi\right)^2$$

satisfies that

i) $\left(\mathcal{F}_c^2(\mathbb{R}^p), D_\theta^\varphi\right)$ is a separable L^2-type metric space.

ii) The support function $s : \mathcal{F}_c^2(\mathbb{R}^p) \to \mathcal{H}$ states an isometrical embedding of $\mathcal{F}_c^2(\mathbb{R}^p)$ onto a closed convex cone of $\mathcal{H}$.

As a result, data in the fuzzy setting with the fuzzy arithmetic and the metric D_θ^φ can be systematically translated into data in the setting of functional values with the functional arithmetic and the metric based on the norm $\|\cdot\|_\theta^\varphi$. In this way, although fuzzy data should not be treated directly as functional data, they can be treated as functional data by considering the identification with their support functions. Many developments in functional data analysis could be applied to fuzzy data by using the appropriate identifications and correspondences, whenever it can be guaranteed that the elements which should belong to $s\left(\mathcal{F}_c^2(\mathbb{R}^p)\right)$ are well-defined within it.

The D_θ^φ metric on $\mathcal{F}_c^2(\mathbb{R}^p)$ can be equivalently expressed as follows

$$D_\theta^\varphi(\widetilde{U}, \widetilde{V}) = \sqrt{\left(\|\text{mid}\,\widetilde{U} - \text{mid}\,\widetilde{V}\|_1^\varphi\right)^2 + \theta\left(\|\text{spr}\,\widetilde{U} - \text{spr}\,\widetilde{V}\|_1^\varphi\right)^2}.$$

For each level, the choice of θ allows us to weight the effect of the deviation between spreads (which can be intuitively translated into the difference in 'shape' or 'imprecision'), in contrast to the effect of the deviation between mid's (intuitively translated into the difference in 'location'). On the other hand, φ has no stochastic but a weighting mission, the choice of φ enables to weight the relevance of different levels.

1.2.5 *Computing the distance in $\mathcal{F}_c^2(\mathbb{R})$*

Given two fuzzy numbers $\widetilde{U}$ and $\widetilde{V}$, the D_θ^φ distance between them is given by the value

$$D_\theta^\varphi(\widetilde{U},\widetilde{V}) = \sqrt{\int_{[0,1]} \left([\mathrm{mid}\, \widetilde{U}_\alpha - \mathrm{mid}\, \widetilde{V}_\alpha]^2 + \theta \cdot [\mathrm{spr}\, \widetilde{U}_\alpha - \mathrm{spr}\, \widetilde{V}_\alpha]^2 \right) \mathrm{d}\varphi(\alpha)}$$

Steps

*S*1.) A 'bijection' between $\widetilde{U}$ and $\widetilde{V}$ is first considered by associating for any arbitrary $\alpha \in [0,1] : \widetilde{U}_\alpha \leftrightarrow \widetilde{V}_\alpha$

*S*2.) The squared of the Euclidean distance between the mid-points is computed

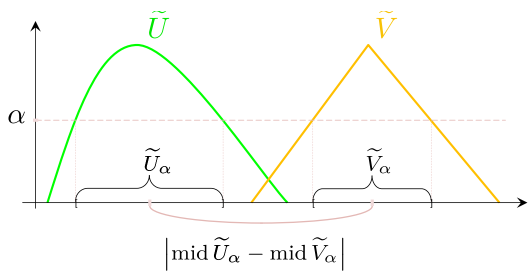

Figure 1.7: Computing the distance in $\mathcal{F}_c^2(\mathbb{R})$ I/II

*S*3.) The squared of the Euclidean distance between the spreads is computed

*S*4.) Two weighted averages, one for each of these squared distances, are considered over different levels, and finally, a weight is assigned to the second one and the sum is computed

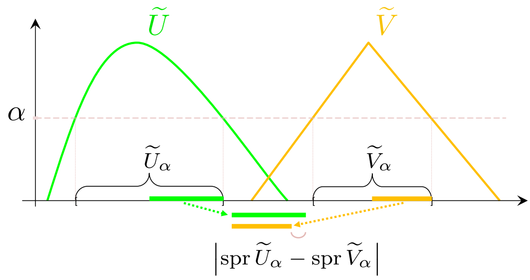

Figure 1.8: Computing the distance in $\mathcal{F}_c^2(\mathbb{R})$ II/II

In the context of this work and the particular case of the considered experiment a $\mathcal{O}(1)$ functional expression of the $\{0,1\}$-levels values is computed, see more details in 5.3.4.

1.3 *Random fuzzy sets and relevant parameters*

Random fuzzy sets (for short RFS) were introduced by Puri and Ralescu[12] , as a mathematical model associating a fuzzy value with each outcome of a random experiment and extending level-wise the concept of random set. They often referred to in the literature as *fuzzy random variables in Puri and Ralescu's sense*). Several measurability conditions are deduced to be equivalent. Namely

[12] M.L. Puri and D.A. Ralescu. Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, 114(2):409–422, 1986

Theorem 1.3.1 *Given a probability space $(\Omega, \mathcal{A}, P)$, consider the mapping $\mathcal{X} : \Omega \to \mathcal{F}_c^2(\mathbb{R}^p)$. Then, the following statements are equivalent:*

i) *$\mathcal{X}$ is a RFS, that is, for all $\alpha \in (0,1]$ the α-level set-valued mapping $\mathcal{X}_\alpha : \Omega \to \mathcal{K}_c(\mathbb{R}^p) = \{\text{nonempty compact convex sets of } \mathbb{R}^p\}$, $\omega \mapsto (\mathcal{X}(\omega))_\alpha$, is a compact convex random set (that is, $\mathcal{A}|\odot-$ measurable, where $\odot$ is the Borel σ-field associated with the Hausdorff metric on $\mathcal{K}_c(\mathbb{R}^p)$).*

ii) *$\mathcal{X}$ is a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by the metric D_θ^φ on $\mathcal{F}_c^2(\mathbb{R}^p)$.*

iii) *$s_\mathcal{X} : \Omega \to \mathcal{H}$ is an $\mathcal{H}$-valued random element, that is, a Borel measurable mapping w.r.t. $\mathcal{A}$ and the Borel σ-field generated by the topology induced by $\|\cdot\|_\theta^\varphi$ on $\mathcal{H}$.*

iv) For all $\alpha \in (0,1]$ and $u \in \mathbb{S}^{p-1}$, the function $s_{\mathcal{X}}(u,\alpha) : \Omega \to \mathbb{R}$ is a real-valued random variable.

v) For all $\alpha \in (0,1]$ and $u \in \mathbb{S}^{p-1}$, the functions $\mathrm{mid}\, s_{\mathcal{X}}(u,\alpha) : \Omega \to \mathbb{R}$ and $\mathrm{spr}\, s_{\mathcal{X}}(u,\alpha) : \Omega \to [0,+\infty)$ are real-valued random variables.

On the basis of Theorem 1.3.1 it is concluded that notions like the *distribution induced by an RFS* or the *stochastic independence of RFS's* are the usual ones for Borel measurable mappings in metric spaces. The mean value of an RFS can be presented in two equivalent ways, either as an extension of the set-valued Aumann expectation or induced from the expectation of an $\mathcal{H}$-valued random element. Thus

Definition 1.3.1 *Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFS $\mathcal{X}$ such that $s_{\mathcal{X}} \in L^1(\Omega, \mathcal{A}, P)$, the (Aumann type)* ***mean value*** *or* ***expected value*** *of $\mathcal{X}$ is the fuzzy value $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R}^p)$ such that for all $\alpha \in (0,1]$*

$$\left(\widetilde{E}(\mathcal{X})\right)_\alpha = \textit{Aumann integral of } \mathcal{X}_\alpha$$

$$= \left\{ \int_{\mathbb{R}^p} X(\omega)\, dP(\omega) \text{ for all } X : \Omega \to \mathbb{R}^p, X \in L^1(\Omega, \mathcal{A}, P), X \in \mathcal{X}_\alpha \text{ a.s. } [P] \right\}$$

or, equivalently, such that

$$s_{\widetilde{E}(\mathcal{X})} = E(s_{\mathcal{X}}).$$

In case $p = 1$, if $\mathcal{X}$ is an RFS such that $\max\{|\inf \mathcal{X}_0|, |\sup \mathcal{X}_0|\} \in L^1(\Omega, \mathcal{A}, P)$, we have that for each $\alpha \in [0,1]$

$$\left(\widetilde{E}(\mathcal{X})\right)_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)].$$

This definition for the mean value is coherent with the considered arithmetic and satisfies the usual properties of linearity. Moreover, it is the Fréchet's expectation w.r.t. D_θ^φ.

The variance of an RFS will be based on the Fréchet's approach. The (θ, φ)-Fréchet variance is conceived as a measure of the error in approximating or estimating the values of the RFS through the corresponding mean value. The real-valued quantification of the dispersion will enable to compare random elements, populations, samples, estimators, etc. by simply ranking real numbers. Due to the properties of the support function and the Hilbertian random elements, the considered variance satisfies the usual properties for this concept.

Definition 1.3.2 *Given a probability space $(\Omega, \mathcal{A}, P)$ and an associated RFS $\mathcal{X}$ such that $s_{\mathcal{X}} \in L^2(\Omega, \mathcal{A}, P)$, the (θ, φ)-**Fréchet variance** of $\mathcal{X}$ is the real number*

$$\sigma^2_{\mathcal{X}} = E\left(\left[D^{\varphi}_{\theta}\left(\mathcal{X}, \widetilde{E}(\mathcal{X})\right]\right)\right]^2\right)$$

or, equivalently,

$$\sigma^2_{\mathcal{X}} = E\left(\left[\left\|s_{\mathcal{X}} - s_{\widetilde{E}(\mathcal{X})}\right\|^{\varphi}_{\theta}\right]^2\right) = E\left(\left[\|s_{\mathcal{X}} - E\left(s_{\mathcal{X}}\right)\|^{\varphi}_{\theta}\right]^2\right) = \mathrm{Var}(s_{\mathcal{X}})$$

$$= E\left(\left[\|\mathrm{mid}\,\mathcal{X} - E\left(\mathrm{mid}\,\mathcal{X}\right)\|^{\varphi}\right]^2\right) + \theta\, E\left(\left[\|\mathrm{spr}\,\mathcal{X} - E\left(\mathrm{spr}\,\mathcal{X}\right)\|^{\varphi}\right]^2\right)$$

$$= \mathrm{Var}(\mathrm{mid}\,\mathcal{X}) + \theta\,\mathrm{Var}(\mathrm{spr}\,\mathcal{X}).$$

2

Classification of Fuzzy Convex Data

2.1 *Previous research*

Colubi et al[1]. tackled this issue on which this work presents some contributions, the starting point is to assume that we have a probability space $(\Omega, \mathcal{A}, P)$, and for each individual we observe a fuzzy datum. Each individual may belong to one of k different categories $g_1, \ldots, g_k$. As learning sample we have n independent individuals, the corresponding fuzzy data and categories. The goal is to find a rule allowing us to assign each new individual one of the k categories.

[1] A. Colubi, G. González-Rodríguez, M. Ángeles Gil, and W. Trutschnig. Nonparametric criteria for supervised classification of fuzzy data. *International Journal of Approximate Reasoning*, 2011

Density-based Classification Criteria for Fuzzy Data (DCCF) : consist in to nonparametrically estimate[2] $P(G = g|\mathcal{X} = \widetilde{x})$, for $g = 1, \ldots, k$, $\widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p)$, and then to assign the new data to the class of maximum estimated probability. That leads to estimate the membership probabilities $p_g = P(G = g|\mathcal{X} = \widetilde{x})$ by means of kernel density estimators and that leads to the necessity of choosing and adequate bandwidth.

[2] Thus assuming the existence of the conditional densities.

Ball-based Classification Criteria for Fuzzy Data (BCCF) : instead of focusing the classification technique on estimating the conditional probabilities $\{P(G = g|\mathcal{X} = \widetilde{x})\}_{g=1}^{k}$ with $\widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p)$, they suggest a classification based on the quantities $\{P(G = g|\mathcal{X} \in B(\widetilde{x}; \delta))\}_{g=1}^{k}$

The BCCF leads to a simpler and more versatile approach, with similar if not better results than the DCCF methods, this work is going to extend the algorithms presented under the BCCF paradigm and do a comparison with traditional supervised classifiers.

2.1.1 *Ball-based Classification Criteria for Fuzzy Data (BCCF)*

The starting point is to try to do the classification based on the quantities

$$\{P(G = g|\mathcal{X} \in B(\widetilde{x};\delta))\}_{g=1}^{k}$$

for a value $\delta > 0$ to be chosen. On the one hand, this simplification could entail a loss of accuracy. On the other hand, it can be applied in a more general setting than DCCF, because no assumption about the existence of the conditional densities is made. Note that if all the conditional densities exist, then BCCF may be formally expressed as a special case of DCCF with a uniform kernel on [0,1] and $h = \delta$.

Assuming $P(X \in B(\widetilde{x};\delta)) > 0$ Bayes Theorem implies

$$P(G = g|\mathcal{X} \in B(\widetilde{x};\delta)) = \frac{P(D_\theta^\varphi(\mathcal{X},\widetilde{x}) \leq \delta|G = g))P(G = g)}{\sum\limits_{l=1}^{k} P(D_\theta^\varphi(\mathcal{X},\widetilde{x}) \leq \delta|G = l)P(G = l)}.$$

Consider $\delta > 0$ and $g \in \{1,\ldots,k\}$. A natural estimator for

$$P(D_\theta^\varphi(\mathcal{X},\widetilde{x}) \leq \delta|G = g)$$

based on the sample information is

$$\widehat{P}(D_\theta^\varphi(\mathcal{X},\widetilde{x}) \leq \delta|G = g) = \frac{n_{\delta,g}}{n_g},$$

where $n_{\delta,g}$ is the number of observations in the sample belonging to group g and for which the distance to $\widetilde{x}$ is lower than or equal to δ. Thus applying Bayes formula,

$P(G = g|\mathcal{X} \in B(\widetilde{x};\delta))$ can nonparametrically be estimated as follows

$$\widehat{P}(G = g|\mathcal{X} \in B(\widetilde{x};\delta)) = \frac{\frac{n_{\delta,g}}{n_g}\frac{n_g}{n}}{\sum_{l=1}^{k}\frac{n_{\delta,l}}{n_l}\frac{n_l}{n}} = \frac{n_{\delta,g}}{\sum_{l=1}^{k} n_{\delta,l}}.$$

To summarize, consider a training sample $\{(\mathcal{X}_i, G_i)\}_{i\in\{1,\ldots,n\}}$ with $\mathcal{X}_i$ a fuzzy datum (in $\mathcal{F}_c(\mathbb{R}^p)$) and G_i the corresponding membership group (in $\{1,\ldots,k\}$). For each $g \in \{1,\ldots,k\}$ let n_g be the number of observations for which $G_i = g$ and denote by $\{\mathcal{Y}_{j,g}\}_{j=1}^{n_g}$ the corresponding collection of such fuzzy data (that is, the conditional samples). Given a fuzzy value $\widetilde{x} \in \mathcal{F}_c(\mathbb{R}^p)$, the suggested Ball-based Classification Criteria for Fuzzy Data (BCCF) can be summarized in the following algorithm

BCCF Algorithm

Step 1. Compute the distance between the datum $\widetilde{x}$ to be classified and the set of training fuzzy data, that is

$$d_{j,g} = D_\theta^\varphi(\widetilde{x}, \mathcal{Y}_{j,g}) \quad \text{for all } j \in 1,\ldots,n_g \text{ and all } g \in \{1,\ldots,k\}.$$

Step 2. Fix a value for $\delta > 0$ and for each $g \in \{1,\ldots,k\}$ compute

$$n_{\delta,g} = \sum_{j=1}^{n_g} I_{[0,\delta]}(d_{j,g}).$$

Step 3. Estimate the membership probabilities $p_g = P(G = g|\mathcal{X} \in B(\widetilde{x};\delta))$ by means of

$$\widehat{P}(G = g|\mathcal{X} \in B(\widetilde{x};\delta)) = \frac{n_{\delta,g}}{\sum_{l=1}^{k} n_{\delta,l}}.$$

Step 4. Assign $\widetilde{x}$ to the group $g(\widetilde{x}) \in \{1,\ldots,k\}$ of maximum estimated probability.

One of the essential issues of this approach is to select a suitable δ. The authors present two approaches for calculating δ

BCCF1: In this first case, δ was chosen to be the maximum of the sample deviations[3] in each group (trying to preserve the simplicity of this method). The reason for considering the maximum instead, for instance, the minimum is to try to

[3] With respect to the central and dispersion measures of definitions 1.3.1 and 1.3.2.

ensure that the balls will be large enough for containing data points of at least one group.

BCCF2: In contrast to the above-mentioned simple approach, a more elaborate selection was considered. Namely δ was chosen as the value maximizing an accuracy measure, concretely the 10-random-3-fold (within each group) classification accuracy.

2.2 *Contributions*

We present two contributions which are enhancements of the BCCF algorithms[4], one is instead of searching for the best δ search for the best θ , in order to simplify the search in the parameter space $\theta \in (0, +\infty)$ we make the inner product defined in 1.2.3 convex, thus

[4] This pattern for optimizing an algorithm is called *Grid Search*.

$$\langle f, g \rangle_{\tau}^{\varphi} = (1-\tau)\left[\operatorname{mid} f, \operatorname{mid} g\right]^{\varphi} + \tau\left[\operatorname{spr} f, \operatorname{spr} g\right]^{\varphi},$$

where $\tau \in (0, 1)$.

This alternative of inner product definition leads to another metric D_{τ}^{φ}, albeit consistent one

$$D_{\tau}^{\varphi}(\widetilde{U}, \widetilde{V}) = \sqrt{(1-\tau)\left(\|\operatorname{mid}\widetilde{U} - \operatorname{mid}\widetilde{V}\|_1^{\varphi}\right)^2 + \tau\left(\|\operatorname{spr}\widetilde{U} - \operatorname{spr}\widetilde{V}\|_1^{\varphi}\right)^2}$$

and the other is a combination of the BFFC2 algorithm and this new algorithm, thus we have

BCCF3: Performs a search in the parameter space τ of the distance (inner product) definition, and chooses the a value maximizing and accuracy measure concretely the 10-random-10-fold (within each group) classification accuracy.

BCCF4: It searches in the $(0, 1)_{\tau} \times [0, \delta_{max}]_{\delta}$ bidimensional space and selects a (τ_i, δ_j) maximizing an accuracy measure, 10-random-10-fold (within each group) classification accuracy.

	τ	δ
BCCF1	fixed	fixed
BCCF2	fixed	search
BCCF3	search	fixed
BCCF4	search	search

Table 2.1: BCCF methods characteristics

We implement[5] all the algorithms including the previously defined BCCF1, BCCF2 using in all cases 10-random 10-fold cross validation at each level of decision, i.e. , for a given (τ_i, δ_j), its performance is evaluated averaging 10 executions at 10 random 10 fold cross validation for each execution.

[5] see 5.4.3 for a more detailed explanation

3
A Visual Perceptions Experiment

The main concepts and methods are to be illustrated by means of a case study which is now introduced along with some guidelines to describe fuzzy data. The case-study regards an experiment in which people have been asked for their perception of the relative length of different line segments with respect to a fixed longer segment that is used as a standard for comparison. Figure 3.1 displays the screen of the application. On the center top of the screen the pattern (longest line segment) is drawn in pink. At each trial a black shorter line segment is generated and placed below the pattern one, parallelly and without considering a concrete location (i.e. indenting or centering).

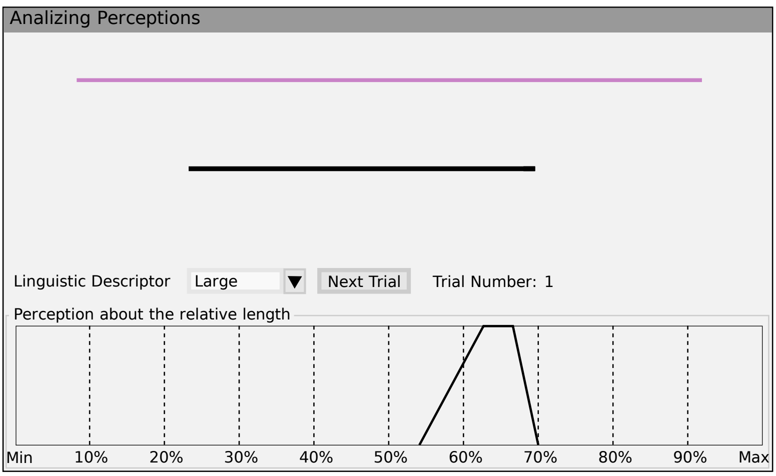

Figure 3.1: SMIRE application to express the perception on the relative length of segments.

Table 3.1 contains some of the data a person making 551 trials delivered[1].

Trial	inf P_0	inf P_1	sup P_1	sup P_0	Ling. descrip.
1	78.27	80.94	84.41	87.40	large
2	54.93	58.00	62.20	65.67	large
3	47.25	49.43	50.89	53.31	medium
4	92.65	95.72	97.58	99.11	very large
5	12.92	15.51	17.77	20.03	very small
6	32.55	36.03	39.90	42.89	small
7	2.50	4.44	6.22	9.21	very small
8	24.80	28.19	30.45	33.28	small
9	55.17	58.40	61.79	65.75	large
10	2.26	3.63	5.57	8.08	very small

Table 3.1: Perceptions about the relative length of the light line segment.

[1] The complete dataset, as well as the software *Perceptions* providing it, can be found at `http://bellman.ciencias.uniovi.es/SMIRE/Perceptions.html`, web page of the SMIRE research group.

After an explanation of the fuzzy values, participants are asked by their judgment of relative length for each of several line segments in two ways. First, to choose a label from a Lickert-like list, {VERY SMALL, SMALL, MEDIUM, LARGE, VERY LARGE}. Second, to describe the perception through a trapezoidal fuzzy number with support included in $[0, 100]$ (0% indicating the minimum relative length and 100 maximum the one). The support is to be chosen as the set of all values that the participant subjectively considers to be compatible with the relative length of the generated segment to a greater or lesser extent. The 1-level has to be the set of all values that the participant considers to be completely compatible with his/her perception about the relative length of the generated segment. The trapezoidal fuzzy set is formed by the linear interpolation of both intervals, although it is possible to change the shape. In other words, out of the support are the values that the participant is not willing to accept as possible values for the relative length at all. The membership degree is linearly increasing from the minimum of the support to the first value for which the participant would say that it is the relative length of the line (see Figure 3.1). However, since the participant may have doubts, often there is not a unique value in these conditions, but an interval. This interval with full membership degree is the 1-level set. Analogously, from the last value in this set and the maximum of the support, the membership degree is linearly decreasing.

The line shown at each trial has been chosen at random - in order to obtain a good coverage of some interesting situations the precise generation procedure was the following one

- 479 lengths were generated from a uniform distribution on $[0, 100]$.
- 9 lengths in the equally spaced discrete set $\{100/27 + i/8\, 100\, (1 - 2/27)\}_{i=0,\ldots,8}$ were repeated 3 times. Thus, we had 27 lengths that are representative of quite different situations that may arise.
- All the random lengths were swapped and shown at random.

In previous research Colubi et al[2]. focused on the problem of finding a rule allowing us to assign each new individual one of the 5 Lickert categories. In Chapter 5 we are also review this problem and present a new one over this dataset, the classification problem of the *Sex* variable of the respondent which records the gender {MALE, FEMALE} of the respondent.

[2] A. Colubi, G. González-Rodríguez, M. Ángeles Gil, and W. Trutschnig. Nonparametric criteria for supervised classification of fuzzy data. *International Journal of Approximate Reasoning*, 2011

4
Fuzzy Data Treated as Compositional Data

4.1 Introduction and motivation

The fuzzy data created by means of the visual perceptions experiment, has besides the convexity property, the constrained property i.e. the last value of the trapezoid must be below 100 and the first value must not be bellow 0.

Given $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ with $\widetilde{U}$ being trapezoidal constrained fuzzy number

$\widetilde{U} = \{\inf \widetilde{U}_0, \inf \widetilde{U}_1, \sup \widetilde{U}_1, \sup \widetilde{U}_0\}$ the following bijection could be established

$$y \in \mathcal{S}^5 = \{\inf \widetilde{U}_0 - 0, \inf \widetilde{U}_1 - \inf \widetilde{U}_0, \sup \widetilde{U}_1 - \inf \widetilde{U}_1, \sup \widetilde{U}_0 - \sup \widetilde{U}_1, 100 - \sup \widetilde{U}_0\}$$

The theory of compositional data is fragile when one of the parts is exactly 0% requiring another treatment. We have 26 degenerate instances where $\inf 0$ is 0. Whenever we apply one of the log ratio transformations on one of those instances we perform a 0.01 translation of the data. Since the data viewed as a compositional data reflects (part) of it's constrained nature it could be of interest to test what kind of results we could get under this framework. The idea is to transform the data to this space, and in this space perform the inference analysis with the tools available[1].

cf. the chapter 'Rounded zeros: some practical aspects for compositional data' in Buccianti et al. for more specific techniques when dealing with this issue.

A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn. Compositional data analysis in the geosciences. *Geological Society Special Publication*, 204, 2007

[1] the log-ratio transformations

4.2 Compositional data fundamentals

Compositional or closed data are multivariate data with positive values that sum up to a constant k , usually chosen as 1 or 100, i.e. $x = \{(x_1, x_2, \ldots, x_D) : x_1 > 0, \ldots, x_D > 0; \sum_{i=1}^{D} x_i = k\}$

The set of all closed observations, denoted as $\mathcal{S}^D$, forms a simplex sample space, a subset of $\mathbb{R}^D$. Standard statistical methods can lead to questionable results if they are directly applied to the original, closed data. For this reason, the family of log-ratio one-to-one transformations from $\mathcal{S}^D$ to the real space was introduced[2]. We will briefly review the basic theory of compositional data that we are going to use in the context of this work.

[2] J. Aitchison. *The statistical analysis of compositional data (2003 reprint)*. Blackburn Press, 1986

- For $D = 3$ the simplex usually is represented in a ternary diagram, an equilateral triangle with k height.
- For $D = 2$ the simplex is represented as a segment.
- Para $D = 4$ the simplex is represented as a tetrahedron.

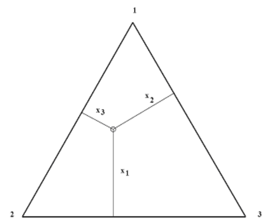

Figure 4.1: Representation of a compositional vector $x = (x_1, x_2, x_3)$ in $\mathcal{S}^3$

4.3 Arithmetic

4.3.1 Closure operator

C is a transformation mapping each vector $w = (w_1, w_2, \ldots, w_D)$ of $\mathbb{R}_+^D$ to its corresponding associated compositional data:

$$C(w) = \left(\frac{k \cdot w_1}{\sum_{i=1}^{D} w_i}, \frac{k \cdot w_2}{\sum_{i=1}^{D} w_i}, \ldots, \frac{k \cdot w_D}{\sum_{i=1}^{D} w_i} \right)$$

4.3.2 Perturbation

The internal operation $\oplus$ perturbation, of one composition x by another composition y refers to the operation,

$x \oplus y = C(x_1 \cdot y_1, x_2 \cdot y_2, \ldots, x_D \cdot y_D)\ \forall x, y \in \mathcal{S}^D$

$(S^D, \oplus)$ is a commutative group, the neutral element of this operation is the barycenter $e_D = \left(\frac{1}{D}, \frac{1}{D}, \ldots, \frac{1}{D}\right) = C(1, 1, \ldots, 1)$

4.3.3 *Powering*

The external operation $\odot$ powering, of one real number λ by another composition x refers to the operation:
$\lambda \odot x = C\left(x_1^{\lambda}, x_2^{\lambda}, \ldots, x_D^{\lambda}\right) \ \forall \lambda \in \mathbb{R}, x \in \mathcal{S}^D$

4.3.4 *Vector space structure*

The tuple $\left(\mathcal{S}^D, \oplus, \odot\right)$ is a vector space of $D-1$ dimensions. The inverse element in this space is defined as: $x^{-1} = -1 \cdot x$, then $x \oplus x^{-1} = x \oplus (-1 \odot x) = e_D$

Usually $\ominus$ denotes the operation of adding the inverse (difference perturbation): $x \oplus y^{-1} = x \oplus (-1 \odot y) = x \ominus y$

4.4 *Distance*

4.4.1 *Euclidean vector space structure*

The inner product in $\left(\mathcal{S}^D, \oplus, \odot\right)$ is defined as,
$\langle x, y \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} \quad x, y \in \mathcal{S}^D$
$\left(\mathcal{S}^D, \oplus, \odot, \langle ., . \rangle_a\right)$ is an Euclidean vector space

4.4.2 *Aitchison's norm*

The inner product induces the following norm,
$\|x\|_a^2 = \langle x, x \rangle_a \Rightarrow \|x\|_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j}\right)^2} \quad x, y \in \mathcal{S}^D$

4.4.3 *Aitchison's distance*

Aichison's norm induces the following metric,

$$d_a^2(x, y) = \|x \ominus y\|_a^2 \Rightarrow d_a(x, y)_a = \sqrt{\frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j}\right)^2} \quad x, y \in \mathcal{S}^D$$

The norm, $\|x\|_a$ can also be seen as the distance from x to the linear space origin, the barycenter: $\|x\|_a = d_a(x, e_D)$

Real space: $\mathbb{R}^{D-1}$	Simplex space: $\mathcal{S}^D$
addition: $x+y$ product: $\alpha \cdot x$ Euclidean distance: $d_e(x,y)$	perturbation: $x \oplus y$ powering: $\alpha \odot x$ Aitchison's distance : $d_a(x,y)$
Vector of means: $\bar{x} = \frac{1}{n} \cdot \sum_{l=1}^{n} x_l$	Metric center: $x = \frac{1}{n} \cdot \bigoplus_{l=1}^{n} x_l = C(g_1, g_2, \ldots, g_D)$ $g_i = (\prod_{l=1}^{n} x_{il})^{\frac{1}{n}} \quad i = 1, \ldots, D$
Distance and translation: $d_e(x+z, y+z) = d_e(x,y)$ Distance and scaling: $d_e(\alpha \cdot x, \alpha \cdot y) = \|\alpha\| d_e(x,y)$	Distance and perturbation: $d_a(x \oplus z, y \oplus z) = d_a(x,y)$ Distance and powering: $d_a(\alpha \odot x, \alpha \odot y) = \|\alpha\| d_a(x,y)$

Table 4.1: Analogy between the real space and the simplex space

4.5 *Simplex transformations*

4.5.1 *Additive log ratio transformation* (alr)

This is a transformation from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$, and the result for an observation $x \in \mathcal{S}^D$ are the transformed data $y \in \mathbb{R}^{D-1}$

$y = (y_1, y_2, \ldots, y_{D-1}) = (\ln \frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D})$

The alr transformation is an isomorphism, but not an isometry $d_a(x,y) \neq d_e(\text{alr}(x), \text{alr}(y))$

4.5.2 *Centered log ratio transformation* (clr)

Compositions $x \in \mathcal{S}^D$ are transformed to data $y \in \mathbb{R}^D$, with $y = (y_1, y_2, \ldots, y_D) = (\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ln \frac{x_2}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}})$

It has the desirable isometry property $d_a(x,y) = d_e(\text{clr}(x), \text{clr}(y))$ but it is easy to see that this transformation results in collinear data because $\sum_{i=1}^{D} yi = 0$. On the other hand, the clr transformation treats all components symmetrically by dividing by the geometric mean. The interpretation of the resulting values might thus be easier.

4.5.3 *Isometric log ratio transformation* (ilr)

This transformation solves the problem of data collinearity resulting from the clr transformation, while preserving all its advantageous properties[3]. It is based on the choice of an orthonormal basis on the hyperplane in $\mathbb{R}^D$ that is formed by the clr transformation, so that the compositions $x \in \mathcal{S}^D$ result in noncollinear data $y \in \mathbb{R}^{D-1}$. The explicit transformation formulas for one such chosen basis are,

$$y = (y_1, y_2, \ldots, y_{D-1}) = (\tfrac{1}{\sqrt{2}} \ln \tfrac{x_1}{x_2}, \tfrac{1}{\sqrt{6}} \ln \tfrac{x_1 x_2}{x_3 x_3}, \ldots, \tfrac{1}{\sqrt{D(D-1)}} \ln \tfrac{\prod_{i=1}^{D-1} x_i}{x_D^{D-1}})$$

Is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, thus avoiding the drawbacks of both the alr and the clr. It has the desirable property $d_a(x,y) = d_e(\text{ilr}(x), \text{ilr}(y))$ but the resulting values are difficult to interpret.

[3] J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3): 279–300, 2003

4.6 *Mapping between constrained trapezoidal convex numbers and compositional data*

Formally we have a datum $\widetilde{U} = \{\inf \widetilde{U}_0, \inf \widetilde{U}_1, \sup \widetilde{U}_1, \sup \widetilde{U}_0\}$,

$$\mathcal{F}_c(\mathbb{R}) \longrightarrow \left(\mathcal{S}^5, \oplus, \odot, \langle .,. \rangle_a\right) \longrightarrow \mathbb{R}^4(+, ., \langle .,. \rangle_2)$$

$$\widetilde{U} \mapsto \text{comp} \mapsto \text{ilr}$$

$$\text{comp}(\widetilde{U}) = \{\inf \widetilde{U}_0 - 0, \inf \widetilde{U}_1 - \inf \widetilde{U}_0, \sup \widetilde{U}_1 - \inf \widetilde{U}_1, \sup \widetilde{U}_0 - \sup \widetilde{U}_1, 100 - \sup \widetilde{U}_0\}$$

$$\begin{aligned}\text{ilr}(\text{comp}(\widetilde{U})) = \Bigg\{ & \frac{1}{\sqrt{2}} \ln \frac{\inf \widetilde{U}_0}{\inf \widetilde{U}_1 - \inf \widetilde{U}_0}, \frac{1}{\sqrt{6}} \ln \frac{\inf \widetilde{U}_0 (\inf \widetilde{U}_1 - \inf \widetilde{U}_0)}{(\sup \widetilde{U}_1 - \inf \widetilde{U}_1)^2}, \\ & \frac{1}{\sqrt{12}} \ln \frac{\inf \widetilde{U}_0 (\inf \widetilde{U}_1 - \inf \widetilde{U}_0)(\sup \widetilde{U}_1 - \inf \widetilde{U}_1)}{(\sup \widetilde{U}_0 - \sup \widetilde{U}_1)^3}, \\ & \frac{1}{\sqrt{24}} \ln \frac{\inf \widetilde{U}_0 (\inf \widetilde{U}_1 - \inf \widetilde{U}_0)(\sup \widetilde{U}_1 - \inf \widetilde{U}_1)(\sup \widetilde{U}_0 - \sup \widetilde{U}_1)}{(100 - \sup \widetilde{U}_0)^4} \Bigg\}\end{aligned}$$

We have then, the following mathematical tool for analyzing this data. Starting with the 4-part $\{0, 1\}$-levels data, map the data to a 5-part composition. Then (re[4])map that composition (back) to the real hyperspace using one of the log-ratio transformations, and then the Euclidean distance in the real space is an Aitchison distance back in the Simplex which in turn is a Fuzzy (Multivariate) distance for the original perception (multivariate) data in the Fuzzy (Multivariate) context .

[4] If we consider the starting point $\mathbb{R}^4$ instead of $\mathcal{F}_c(\mathbb{R})$

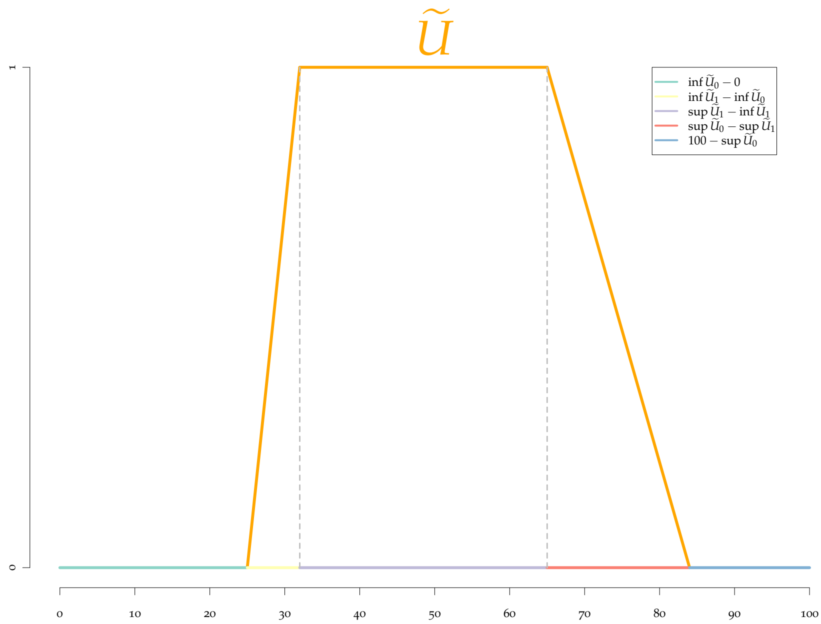

Figure 4.2: Mapping of a 4 part convex constrained fuzzy number to a 5 part composition

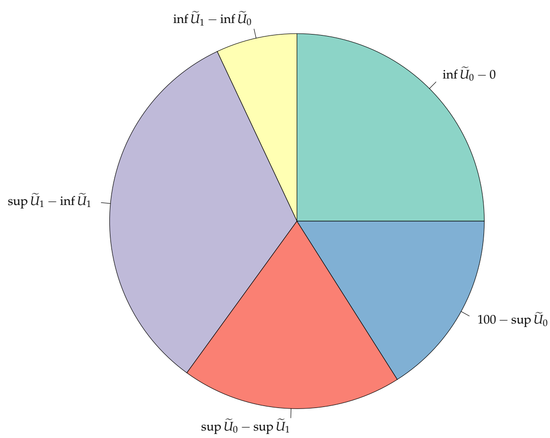

Figure 4.3: 5 part composition of trapezoid convex constrained number of Figure 4.2

5
Classification Analysis

We complement the original supervised classification problem with a new batch of classifiers and present a new problem, the gender of the respondent.

5.1 Dataset Description

The data and software for conducting the experiments are avaible at http://bellman.ciencias.uniovi.es/SMIRE/Perceptions.html
The dataset has following rows:

- EmailId: Identification index for the different emails.
- Sex: Genre of the individual coded as 0=Male, 1=Female.
- nTrial: Number of trial (first 3 blocks of 9 observations, i.e. nTrial from 1 to 27, correspond to random ordered relative sizes in the fixed grid 3.7%, 15.2%, 26.8%, 38.4%, 50.0%, 61.5%, 73.1%, 84.7%, 96.3%, the remaining ones are just selected uniformly at random between 0 and 100).
- pHWS: Number of pixels in the horizontal direction of the Working Space.
- Ref: Relative size of the reference line with respect to the available horizontal working space expressed in percentage (fixed to 80%).
- Pref: Absolute number of pixels (in horizontal sense) occupied by the reference line (i.e. integer part of $\mathrm{Ref} \cdot \mathrm{HWS}/100$).

- Size: Number of pixels (in the horizontal direction) occupied by the line shown.
- Rel.Size: Effective relative size of the line shown with respect to the reference line expressed in percentage (i.e. $100 \cdot \text{Size}/\text{Pref}$). Note that depending on the screen resolution this relative size could be slightly different to the ones selected in the fixed grid for the first 27 trials.
- Ling: Linguistic description selected after fixing the fuzzy response (0=Very Small, 1=Small, 2=Medium, 3=Large, 4=Very Large).
- $\inf 0$: Infima of the 0-cut of the fuzzy set selected.
- $\inf 1$: Infima of the 1-cut of the fuzzy set selected.
- $\sup 1$: Suprema of the 1-cut of the fuzzy set selected.
- $\sup 0$: Suprema of the 0-cut of the fuzzy set selected.
- Date: Date and time when the measurement was done.

In total we have 1387 instances, and 24 different users, the number of replies and genre of each user are shown in Table 5.1.

EmailId	Replies	Sex
1	13	Male
2	50	Male
3	32	Female
3	32	Female
4	30	Female
5	32	Male
6	30	Male
7	31	Female
8	10	Male
9	30	Male
10	31	Female
11	30	Male
12	30	Female
13	41	Female
14	31	Male
15	20	Male
16	31	Female
17	35	Male
18	32	Female
19	30	Male
20	30	Male
21	30	Female
22	551	Male
23	102	Female
24	105	Female

Table 5.1: Number of replies and genre in Perceptions dataset.

Also we have 26 degenerate instances on which $\inf 0$ is 0. That is an issue with the compositional data treatment we are going to present, see Section 4.1 margin note

Males have 892 (64%) replies, and females 495 (36%), the most frequent user is EmailId: 22.

The data $\{\inf 0, \inf 1, \sup 1, \sup 0\}$ is highly correlated.

	inf 0	inf 1	sup 1	sup 0
inf 0	1	0.998	0.990	0.983
inf 1	0.998	1	0.995	0.990
sup 1	0.990	0.995	1	0.998
sup 0	0.983	0.990	0.998	1

Table 5.2: Correlation matrix of the α-cuts.

5.2 *Problems description*

For this work we are going to concentrate on the subset $\{\inf 0, \inf 1, \sup 1, \sup 0, \text{Ling}, \text{Sex}\}$ and try to do inference of

1. *Ling* $\sim \{\inf 0, \inf 1, \sup 1, \sup 0\}$, classify *Ling* given the fuzzy perception, this is going to be referred as the *Ling Classification problem*.

2. *Sex* $\sim \{\inf 0, \inf 1, \sup 1, \sup 0\}$ classify *Sex*, this is going to be referred as the *Sex Classification problem*.

The *Ling* classification problem is going to be presented in two versions, $Ling_{[22]}$ which represents the problem analyzed for the user EmailId: 22, and the full dataset which is going to be referred as *Ling*. The first problem was already analyzed by Colubi et al.[1]

[1] A. Colubi, G. González-Rodríguez, M. Ángeles Gil, and W. Trutschnig. Nonparametric criteria for supervised classification of fuzzy data. *International Journal of Approximate Reasoning*, 2011

5.3 *Methodology*

The inference is going to be made under the optic of different algorithms and different representations of the covariate data.

5.3.1 *Data representation*

- *(Direct) Multivariate space*: treats the data as a $\mathbb{R}^n$ point, with the usual Euclidean arithmetic. Represents the raw data, treated as real multivariate data, and direct combinations of such data In this space we consider the following sets

  $\alpha_{\{0,1\}} = \{\inf 0, \inf 1, \sup 1, \sup 0\}$

  steiner calculated using the SAFD[2] package, which as weighting measure uses the Lebesgue measure on [0,1].

[2] Wolfgang Trutschnig and Asun Lubiano. *SAFD: Statistical Analysis of Fuzzy Data*, 2011. URL http://CRAN.R-project.org/package=SAFD. R package version 0.3

$\text{mid}_{\alpha_1} = \frac{\sup 1 + \inf 1}{2}$

$\text{sprs} = \{\inf 1 - \sup 0, \sup 0 - \sup 1, \sup 1 - \text{mid}\, 1\}$[3]

$\text{mid}_{\alpha_1} \cup \text{sprs} = \{\text{mid}_{\alpha_1}, \inf 1 - \sup 0, \sup 0 - \sup 1, \sup 1 - \text{mid}\, 1\}$

[3] Which are respectively the left and right 0-level spreads, and the 1-level spread

- *Simplex space*: Represents the data viewed as a 5-part compositional point and it's log-ratio transformations In this space we consider the following sets

  $\text{comp} = \{\{\inf 0 - 0, \inf 1 - \inf 0, \sup 1 - \inf 1, \sup 0 - sup 1, 100 - \sup 0\}$
  which is the 5 part composition mapping of the fuzzy convex constrained trapezoid.

  $\text{alr} = \text{alr}\,(\text{comp})$ see 4.5.1.

  $\text{clr} = \text{clr}\,(\text{comp})$ see 4.5.2.

  $\text{ilr} = \text{ilr}\,(\text{comp})$ see 4.5.3.

  The data is treated with the usual Euclidean arithmetic, and in the case of the clr (in $\mathbb{R}^5$), and ilr (in $\mathbb{R}^4$). This Euclidian treatment is equivalent to a $\left(\mathcal{S}^5, \oplus, \odot, \langle ., . \rangle_a\right)$ compositional treatment.

- *Fuzzy space*: considers the 4 point (pseudo) multivariate real data as a 1 point in $\left(\mathcal{F}_c^2(\mathbb{R}), D_\tau^\lambda\right)$.

5.3.2 *Data representation example*

Let $\widetilde{U}$ be a perception given by
$\widetilde{U} = \{\inf 0, \inf 1, \sup 1, \sup 0\} = \{78.27, 80.94, 84.41, 87.4\}$, we have

- *Multivariate space*

  $\alpha_{\{0,1\}}(\widetilde{U}) = \{78.27, 80.94, 84.41, 87.4\}$

  $\text{steiner}(\widetilde{U}) = \{82.755\}$

  $\text{mid}\,\alpha_1(\widetilde{U}) = \{82.675\}$

  $\text{sprs}(\widetilde{U}) = \{2.67, 2.99, 1.735\}$

  $\text{mid}\,\alpha_1 \cup \text{sprs}\,(\widetilde{U}) = \{82.675, 2.67, 2.99, 1.735\}$

- *Simplex space*

$$\mathrm{comp}(\widetilde{U}) = \{78.27, 2.67, 3.47, 2.99, 12.6\}$$
$$\mathrm{alr}(\mathrm{comp}) = \{1.8265, -1.5516, -1.2895, -1.4384\}$$
$$\mathrm{clr}(\mathrm{comp}) = \{2.3171, -1.0610, -0.79892, -0.9478, 0.49062\}$$
$$\mathrm{ilr}(\mathrm{comp}) = \{2.3887, 1.1651, 0.95279, -0.54853\}$$

- *Fuzzy space*

$\widetilde{U} = \{78.27, 80.94, 84.41, 87.4\}$ as a point in $\mathcal{F}_c(\mathbb{R})$

5.3.3 *Supervised classification algorithms*

The idea is to compare the BCCF fuzzy algorithms to classical classifiers.

For the *Ling* variable, as it has an implicit 'metric' due to it's ordered nature, an Ordinal wrapper could be employed in conjunction with the classifiers. In exploratory analysis we obtained no improvement using this technique, so we opted to not pursue further with that path.

- For the Multivariate space and the Simplex space

    Linear Discriminant Analysis (LDA)

    Logistic Regression (LR)

    Support Vector Machines (SVM)

    Neural Network (NN)

    k-Nearest Neighbor (*k*-NN)

    $1 -$ Rules (1R)

    C4.5

- For the Fuzzy space: BCCF1, BCCF2, BCCF3, BCCF4

5.3.4 *BCCF classifiers implementation details*

In order to speed up the computations[4] of the distance between to fuzzy numbers. We exploit the discrete convex nature of the trapezoidal fuzzy numbers in question. Thus

$$\mathrm{mid}\,\widetilde{A}_\alpha = \mathrm{mid}\,\widetilde{A}_0 - \alpha\,(\mathrm{mid}\,\widetilde{A}_0 - \mathrm{mid}\,\widetilde{A}_1)$$
$$\mathrm{spr}\,\widetilde{A}_\alpha = \mathrm{spr}\,\widetilde{A}_0 - \alpha\,(\mathrm{spr}\,\widetilde{A}_0 - \mathrm{spr}\,\widetilde{A}_1)$$

[4] An $\mathcal{O}(1)$ strategy given the fixed number of examples could be to precompute all the possible distances combinations, but that does not work for BCCF3 and BCCF4 given it's parameterizable metric definition.

$$\mathrm{mid}\,A_\alpha = \frac{\sup A_\alpha + \inf A_\alpha}{2}$$
$$\mathrm{spr}\,A_\alpha = \sup A_\alpha - \mathrm{mid}\,A_\alpha$$

Treating each α-level as equally important and therefore use the Lebesgue measure as weighting measure $\varphi = \lambda$, we have

$$\left[D_\tau^\lambda(\widetilde{A}, \widetilde{B})\right]^2 = \int_{[0,1]} (1-\tau)[\mathrm{mid}\,\widetilde{A}_\alpha - \mathrm{mid}\,\widetilde{B}_\alpha]^2 + \tau \cdot [\mathrm{spr}\,\widetilde{A}_\alpha - \mathrm{spr}\,\widetilde{B}_\alpha]^2\,\mathrm{d}\alpha$$

let

$$I1 = \int_{[0,1]} [\text{mid}\,\widetilde{A}_\alpha - \text{mid}\,\widetilde{B}_\alpha]^2\,d\alpha$$
$$I2 = \int_{[0,1]} \cdot [\text{spr}\,\widetilde{A}_\alpha - \text{spr}\,\widetilde{B}_\alpha]^2\,d\alpha$$

we have then

$$\begin{aligned}
I1 &= \int_{[0,1]} \big[\big(\text{mid}\,\widetilde{A}_0 - \alpha\cdot(\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big) - \big(\text{mid}\,\widetilde{B}_0 - \alpha\cdot(\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1)\big)\big]^2 \\
&= \int_{[0,1]} \big[(\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{B}_0) + \alpha\big((\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1) - (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big)\big]^2\,d\alpha \\
&= \int_{[0,1]} (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{B}_0)^2\,d\alpha \\
&\quad + \int_{[0,1]} \alpha^2\big((\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1) - (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big)^2\,d\alpha \\
&\quad + \int_{[0,1]} 2\alpha(\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{B}_0)\big((\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1) - (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big)\,d\alpha \\
&= \Bigg\{(\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{B}_0)^2 + \frac{\big((\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1) - (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big)^2}{3} \\
&\quad + (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{B}_0)\big((\text{mid}\,\widetilde{B}_0 - \text{mid}\,\widetilde{B}_1) - (\text{mid}\,\widetilde{A}_0 - \text{mid}\,\widetilde{A}_1)\big)\Bigg\}
\end{aligned}$$

proceeding in a similar manner for I2

$$\begin{aligned}
I2 &= \Bigg\{(\text{spr}\,\widetilde{A}_0 - \text{spr}\,\widetilde{B}_0)^2 + \frac{\big((\text{spr}\,\widetilde{B}_0 - \text{spr}\,\widetilde{B}_1) - (\text{spr}\,\widetilde{A}_0 - \text{spr}\,\widetilde{A}_1)\big)^2}{3} \\
&\quad + (\text{spr}\,\widetilde{A}_0 - \text{spr}\,\widetilde{B}_0)\big((\text{spr}\,\widetilde{B}_0 - \text{spr}\,\widetilde{B}_1) - (\text{spr}\,\widetilde{A}_0 - \text{spr}\,\widetilde{A}_1)\big)\Bigg\}
\end{aligned}$$

so the distance could be computed as

$$D^\lambda_\tau(\widetilde{A}, \widetilde{B}) = \sqrt{(1-\tau)\,I1 + \tau\,I2}$$

5.4 Experimental setup

5.4.1 Software

The data was prepared with R[5] (2.15.0 version)

LDA was run in R.

The classifiers LR, SVM, NN, *k*-NN, 1R, C4.5 were run in Weka[6] (3.7 version) in Experimenter mode.

BCCF methods were implemented in C++.

All the output from the classifiers was postprocessed in R. 'xtable'[7] R package was used to generate the tables in LaTeX code. To generate boxplots in TikZ graphic format from R 'tikzDevice'[8] was employed.

[5] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0

[6] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009

[7] David B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2012. URL http://CRAN.R-project.org/package=xtable. R package version 1.7-0

5.4.2 Inference Schema

Schematically we have the following

A set of problems: $\mathcal{P} = \{Ling_{[22]}, Ling, Sex\}$

A set of spaces: $\Xi = \{\text{Multivariate}, \text{Simplex}, \text{Fuzzy}\}$

Each space has a set of algorithms:

$\Lambda_{\{\text{Multivariate},\text{Simplex}\}} = \{\text{LDA}, \text{LR}, \text{SVM}, \text{NN}, \text{kNN}, \text{1R}, \text{C4.5}\}$

$\Lambda_{\text{Fuzzy}} = \{\text{BCCF1}, \text{BCCF2}, \text{BCCF3}, \text{BCCF4}\}$

Each space has different ways of presenting the data to the inferred variable

$\Xi_{\text{Multivariate}} = \{\alpha_{\{0,1\}}, \text{steiner}, \text{mid}\,\alpha_1, \text{sprs}, \{\text{mid}\,\alpha_1 \cup \text{sprs}\}\}$

$\Xi_{\text{Simplex}} = \{\text{comp}, \text{alr}\,, \text{clr}\,, \text{ilr}\,\}$

$\Xi_{\text{Fuzzy}} = \{\alpha_{\{0,1\}_{\mathcal{F}_C^D(\mathbb{R})}}\}$

$$\mathcal{P}_k \overset{\Lambda_{\xi_j}}{\sim} \Xi_{\xi_i}$$

Given a problem k, is inferred in a space ξ using a representation i with an algorithm j for the given space.

[8] Charlie Sharpsteen and Cameron Bracken. *tikzDevice: A Device for R Graphics Output in PGF/TikZ Format*, 2012. URL http://CRAN.R-project.org/package=tikzDevice. R package version 0.6.2

5.4.3 Classifier evaluation

In all the problems, we employed a stratified 10-fold cross-validation, and run 100 global executions. For the *Sex* class problem, due to the unbalanced nature of the dataset—64%

males, 36% females—we opted for producing 10 random balanced datasets having all the female instances and a random sample of size females for the males instances. Each of these 10 datasets were evaluated globally 10 times.

Whenever a parameter was inferred for the algorithm, it was done without using the test *k*-fold. Instead a new subproblem was created with the training dataset, which was subdivided in a 10 fold cross validation problem. Thus when evaluating a fold, the test data is not used for adjusting the algorithm.

5.5 $\text{Ling}_{[22]}$

5.5.1 *Linear Discriminant Analysis*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	89.45	90.19	90.11	36.63	89.32
Median	90.38	90.94	90.87	38.46	90.38
Mean	90.37	90.96	90.84	38.46	90.38
Maximun	91.45	91.88	91.51	40.95	91.16
Deviation	0.30	0.26	0.26	0.61	0.29

Table 5.3: Summary of the percentage of correct classification for $\textit{Ling}_{[22]}$ variable with LDA method in Multivariate space.

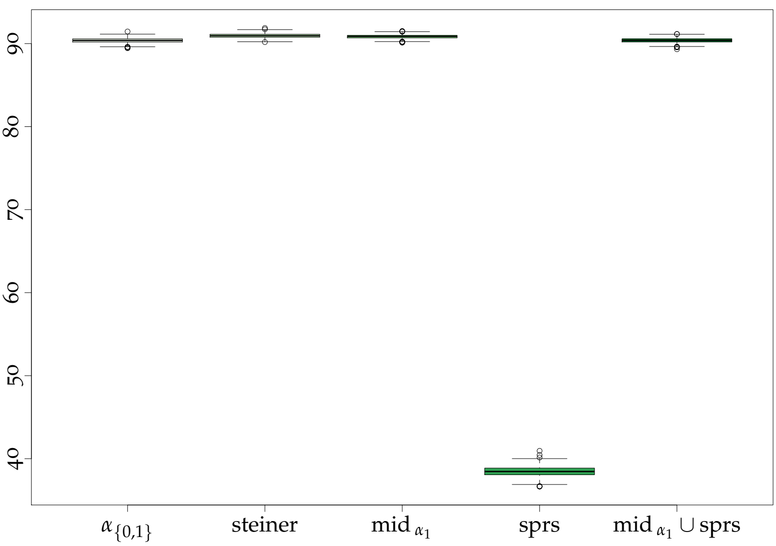

Figure 5.1: Boxplots of the percentage of correct $\textit{Ling}_{[22]}$ classification with LDA method in Multivariate space.

The performance is excellent in all datasets but sprs[9], which contains no information about the relationship between a perception and a linguistic description. The algorithm is slightly more stable in 1-arity datasets. Notice a small degradation of performance in the mid $_{\alpha_1} \cup$ sprs with respect to mid $_{\alpha_1}$.

[9] This fact, which is present in all the remaining experiments, is not going to be repeated for the rest of results of each individual pair algorithm, representation.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	89.44	71.58	71.49	71.49
Median	90.37	73.29	73.27	73.27
Mean	90.37	73.28	73.26	73.26
Maximun	91.32	74.87	75.63	75.63
Deviation	0.30	0.58	0.58	0.58

Table 5.4: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with LDA method in Simplex space.

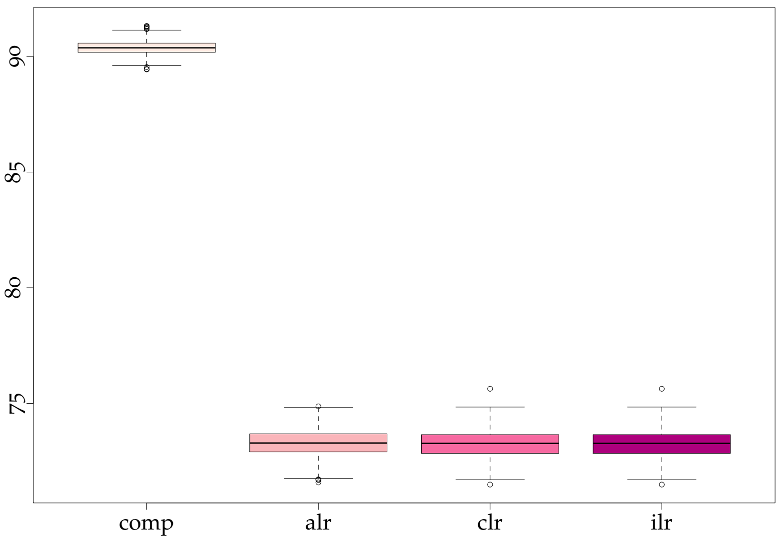

Figure 5.2: Boxplots of the percentage of correct $Ling_{[22]}$ classification with LDA method in Simplex space.

The performance in the comp representation is the same than in the $\alpha_{\{0,1\}}$. In the log-ratio transformations the performance degrades considerably.

5.5.2 *Logistic Regression*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	88.93	90.19	90.37	37.58	88.93
Median	90.20	90.92	90.75	39.03	90.20
Mean	90.24	90.87	90.78	39.08	90.24
Maximun	91.29	91.30	91.30	40.49	91.29
Deviation	0.39	0.20	0.20	0.64	0.39

Table 5.5: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with LR method in Multivariate space.

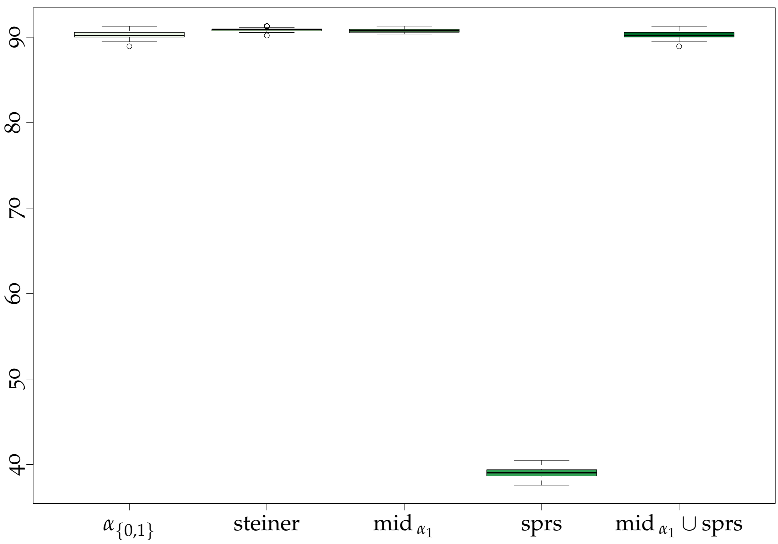

Figure 5.3: Boxplots of the percentage of correct $Ling_{[22]}$ classification with LR method in Multivariate space.

Similar results as LDA, a little more unstable in $\alpha_{\{0,1\}}$ but a bit more stable with steiner and mid $_{\alpha_1}$ datasets. Again there is a degradation of performance in mid $_{\alpha_1} \cup$ sprs with respect to mid $_{\alpha_1}$ and the instability nearly doubles. This kind of argument of degradation in a redundant representation could also be said of the $\alpha_{\{0,1\}}$ with respect to the 1-arity datasets.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	88.93	89.11	89.11	89.11
Median	90.20	90.37	90.37	90.37
Mean	90.24	90.35	90.35	90.35
Maximun	91.29	91.30	91.30	91.30
Deviation	0.39	0.46	0.46	0.46

Table 5.6: Summary of the percentage of correct classification for *Ling*$_{[22]}$ variable with LR method in Simplex space.

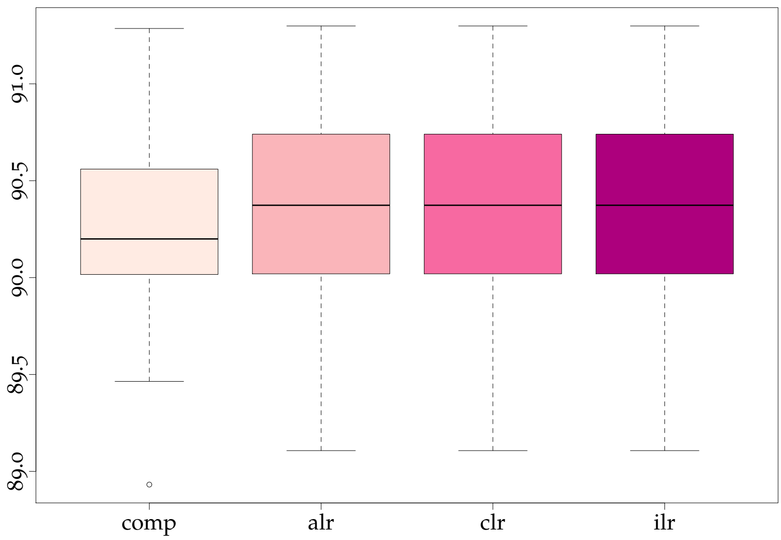

Figure 5.4: Boxplots of the percentage of correct *Ling*$_{[22]}$ classification with LR method in Simplex space.

Similar results as with the Multivariate space in mean value, but the variability doubles for the log-ratio transformations.

5.5.3 *Support Vector Machines*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	87.83	88.93	88.76	37.75	87.30
Median	89.48	90.11	90.01	39.39	88.56
Mean	89.48	90.09	89.92	39.44	88.56
Maximun	90.75	91.11	91.47	40.84	90.20
Deviation	0.58	0.49	0.52	0.69	0.56

Table 5.7: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with SVM method in Multivariate space.

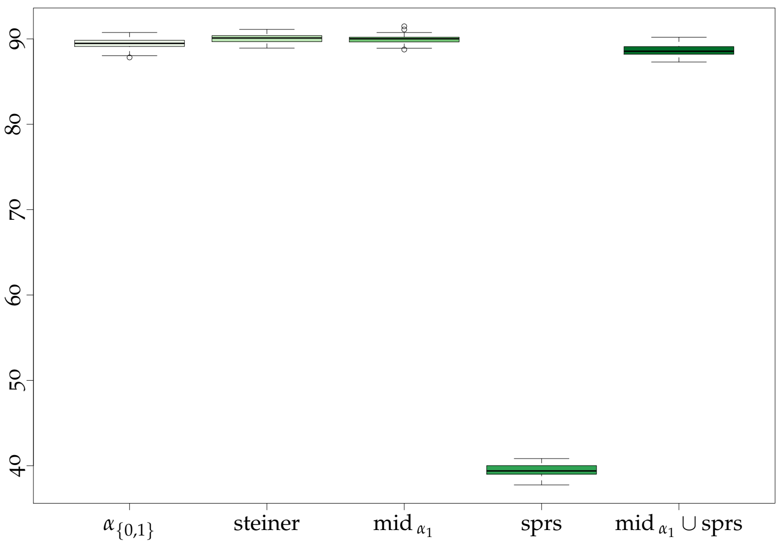

Figure 5.5: Boxplots of the percentage of correct $Ling_{[22]}$ classification with SVM method in Multivariate space.

Excellent results in all datasets with location information, a bit more unstable and a slightly worse mean performance than the simpler LDA, LR.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	88.56	88.56	88.76	88.76
Median	89.29	89.47	89.83	89.83
Mean	89.30	89.43	89.78	89.78
Maximun	90.19	90.03	90.74	90.74
Deviation	0.30	0.30	0.32	0.32

Table 5.8: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with SVM method in Simplex space.

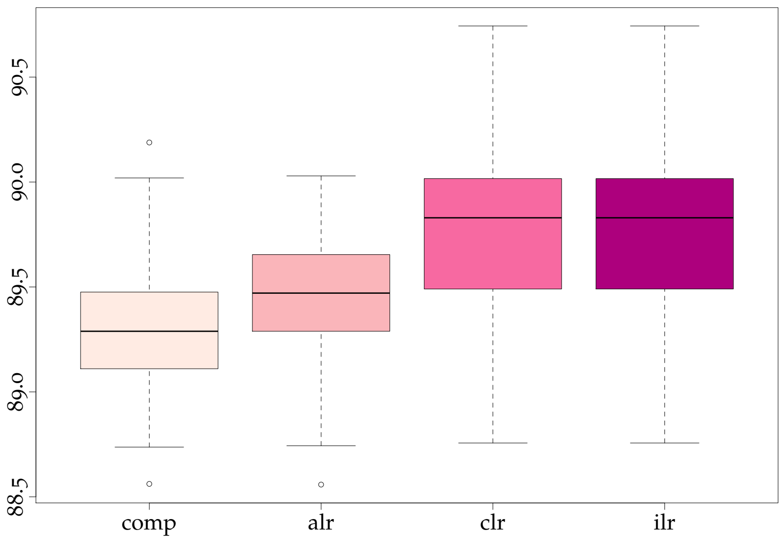

Figure 5.6: Boxplots of the percentage of correct $Ling_{[22]}$ classification with SVM method in Simplex space.

Similar to the multivariate case in mean performance but with much more stability. Notice the gain in stability in the comp representation with respect to $\alpha_{\{0,1\}}$.

5.5.4 *Neural Network*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	86.74	77.84	78.04	35.23	86.20
Median	89.84	83.30	82.95	38.68	88.74
Mean	89.72	83.21	83.06	38.63	88.72
Maximun	91.28	87.83	87.67	41.38	90.92
Deviation	0.84	1.98	1.97	1.18	0.95

Table 5.9: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with NN method in Multivariate space.

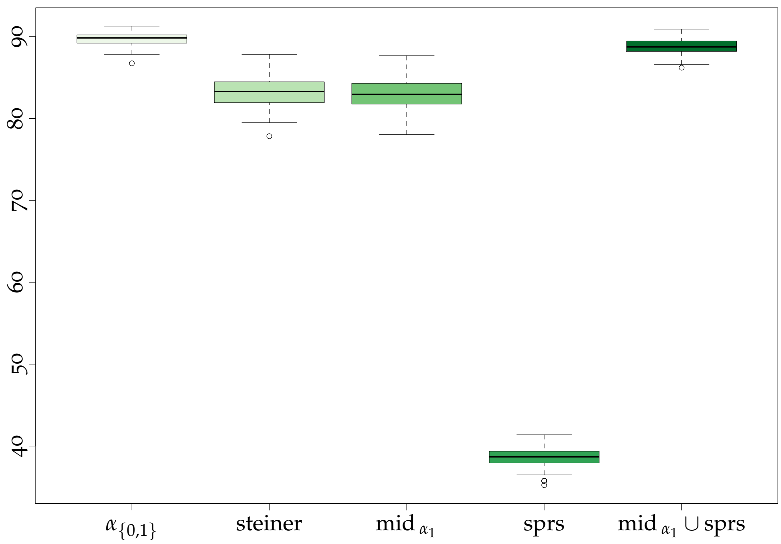

Figure 5.7: Boxplots of the percentage of correct $Ling_{[22]}$ classification with NN method in Multivariate space.

The performance degrades in 1-arity datasets. We have an inversion of the inversion[10]. The $\alpha_{\{0,1\}}$ representation produces better results and more stable than its 1-arity counterparts. The stability is higher than in simpler algorithms with worse performance.

[10] cf. the commentary for the Multivariate space in 5.5.1

– **Simplex space**

	comp	alr	clr	ilr
Minimum	87.47	80.93	82.20	82.20
Median	89.11	86.38	86.57	86.57
Mean	89.18	86.09	86.54	86.54
Maximun	90.93	88.39	88.93	88.93
Deviation	0.82	1.37	1.31	1.31

Table 5.10: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with NN method in Simplex space.

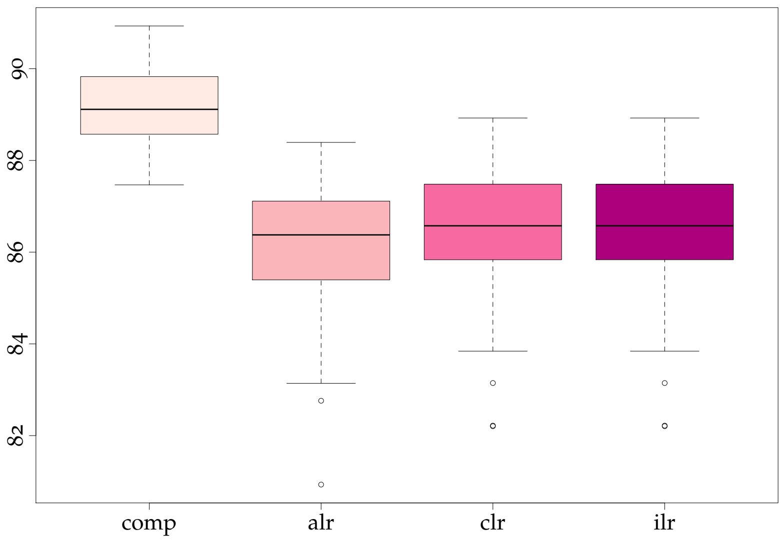

Figure 5.8: Boxplots of the percentage of correct $Ling_{[22]}$ classification with NN method in Simplex space.

The log-ratio transformation offers no improvement with respect to the Multivariate space. The comp representation behaves in a similar way to the $\alpha_{\{0,1\}}$ representation.

5.5.5 k-Nearest Neighbor

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	87.11	88.56	87.46	33.57	83.31
Median	88.92	90.02	89.48	36.12	84.76
Mean	88.94	90.07	89.47	36.13	84.77
Maximun	90.92	91.11	90.39	39.03	86.40
Deviation	0.67	0.56	0.53	1.13	0.67

Table 5.11: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with k-NN method in Multivariate space.

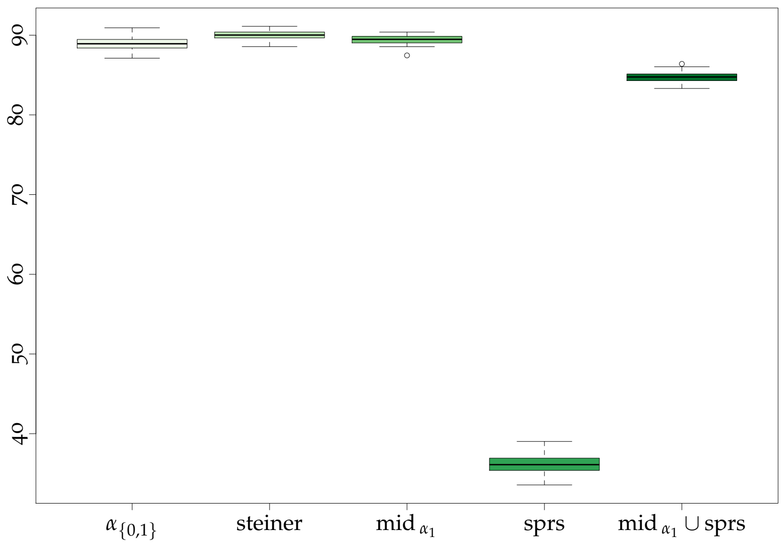

Figure 5.9: Boxplots of the percentage of correct $Ling_{[22]}$ classification with k-NN method in Multivariate space.

Excellent results for $\alpha_{\{0,1\}}$, steiner, and mid $_{\alpha_1}$. The degradation in the redundant dataset mid $_{\alpha_1} \cup$ sprs with respect to mid $_{\alpha_1}$ is about 25% here. [11]

[11] Which clearly points out that spreads for the $Ling_{[22]}$ classification introduce noise. Due to the nature of the method, this noise is squared.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	84.57	85.65	82.76	82.76
Median	86.57	87.47	84.21	84.21
Mean	86.59	87.37	84.17	84.17
Maximun	88.38	89.30	85.48	85.48
Deviation	0.69	0.65	0.61	0.61

Table 5.12: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with k-NN method in Simplex space.

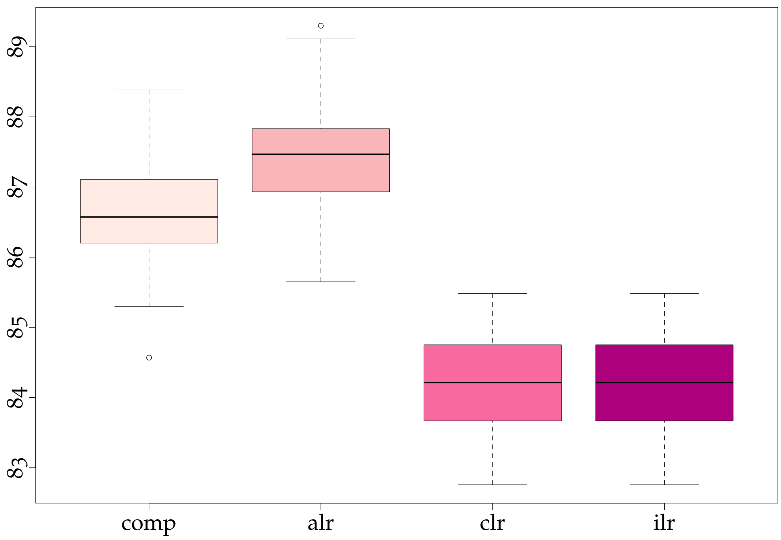

Figure 5.10: Boxplots of the percentage of correct $Ling_{[22]}$ classification with k-NN method in Simplex space.

Slightly worse results than in the simpler Multivariate configurations.

5.5.6 1 – Rules

– Multivariate space

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	89.29	89.65	88.02	26.14	88.02
Median	90.57	90.92	89.12	30.40	89.12
Mean	90.57	90.85	89.09	30.42	89.09
Maximun	91.65	92.03	90.38	34.49	90.38
Deviation	0.56	0.57	0.54	1.51	0.54

Table 5.13: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with 1R method in Multivariate space.

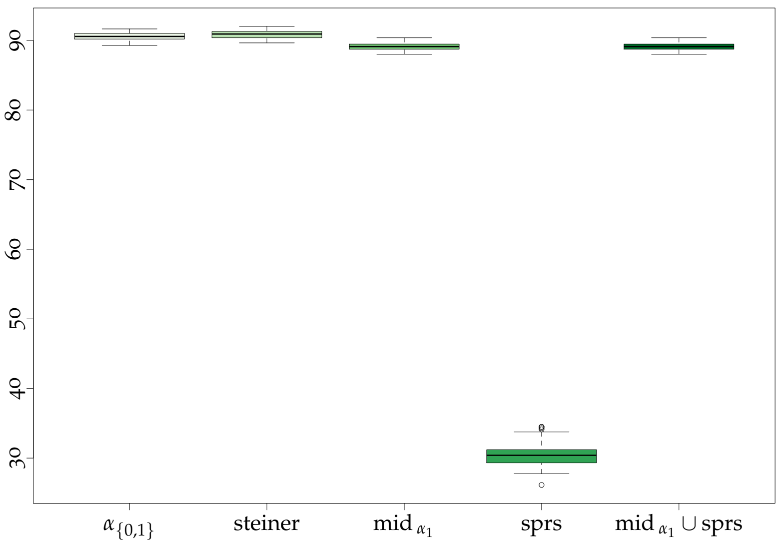

Figure 5.11: Boxplots of the percentage of correct $Ling_{[22]}$ classification with 1R method in Multivariate space.

Excellent performance, better than the more complex NN and on par with SVM. Notice the robustness of the method with respect to noise information. There is no degradation in the mid $_{\alpha_1} \cup$ sprs dataset due to the nature of the method.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	87.48	88.93	74.94	74.94
Median	89.74	90.75	78.40	78.40
Mean	89.69	90.77	78.32	78.32
Maximun	91.12	91.84	80.77	80.77
Deviation	0.73	0.51	1.06	1.06

Table 5.14: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with 1R method in Simplex space.

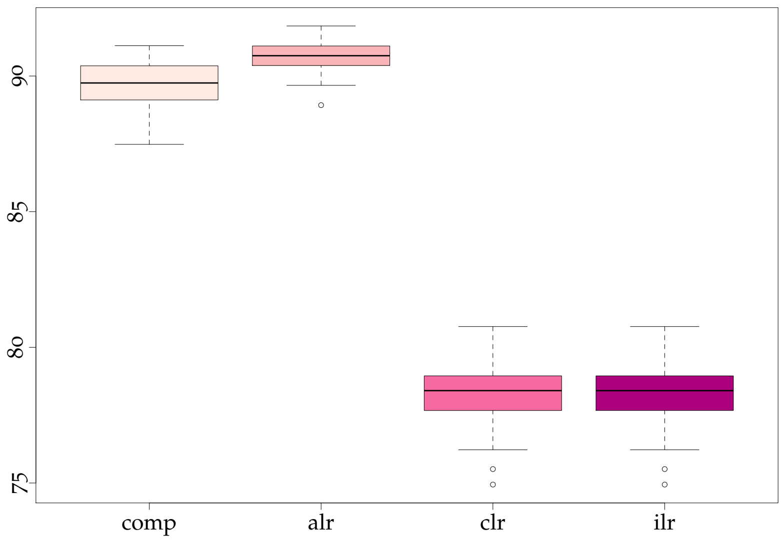

Figure 5.12: Boxplots of the percentage of correct $Ling_{[22]}$ classification with 1R method in Simplex space.

The comp representation behaves slightly worse than the $\alpha_{\{0,1\}}$. The alr transformation performs on par with the non-transformed representations. But the two other ones degrade considerably. The complexity in the clr and ilr transformations introduces noise components in all the resulting transformed variables. And each of the resulting transformed variable carries noise not present in one of the alr coordinates.[12]

[12] The decisive covariate in the alr transformation for 1R is $\ln \frac{\inf 0 - 0}{100 - \sup 0}$, and corresponds to the transformation of the covariate that carries the location information of the original data.

5.5.7 C4.5

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	86.93	88.56	86.92	27.59	86.75
Median	88.39	89.84	88.01	31.57	88.01
Mean	88.37	89.90	88.02	31.62	87.99
Maximun	89.47	90.93	89.29	35.58	89.66
Deviation	0.59	0.54	0.53	1.48	0.64

Table 5.15: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with C4.5 method in Multivariate space.

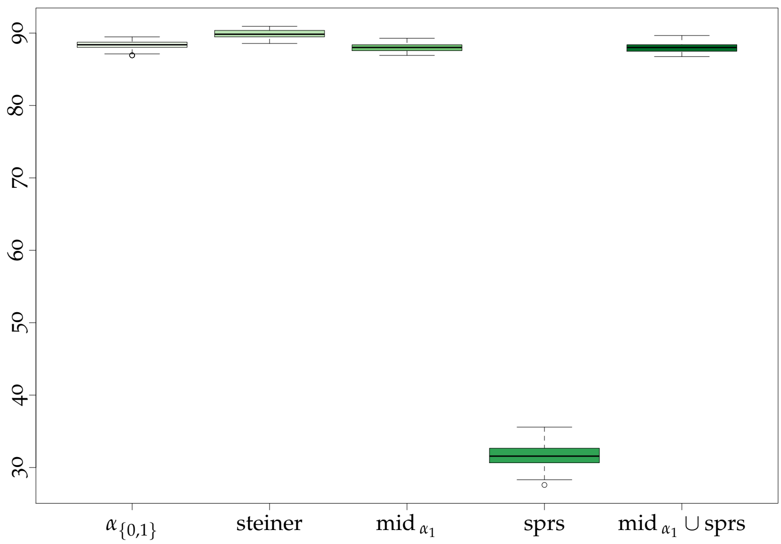

Figure 5.13: Boxplots of the percentage of correct $Ling_{[22]}$ classification with C4.5 method in Multivariate space.

Similar but very slightly worse performance than 1R, which implies that the tree C4.5 is doing splits on noise components or highly redundant[13] data which is the case for the original data $\alpha_{\{0,1\}}$ and its linear combinations.

[13] See the correlation matrix on Table 5.2

– **Simplex space**

	comp	alr	clr	ilr
Minimum	86.38	88.21	85.48	85.48
Median	88.03	89.83	87.30	87.30
Mean	88.13	89.73	87.27	87.27
Maximun	89.30	90.93	89.49	89.49
Deviation	0.64	0.61	0.76	0.76

Table 5.16: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with C4.5 method in Simplex space.

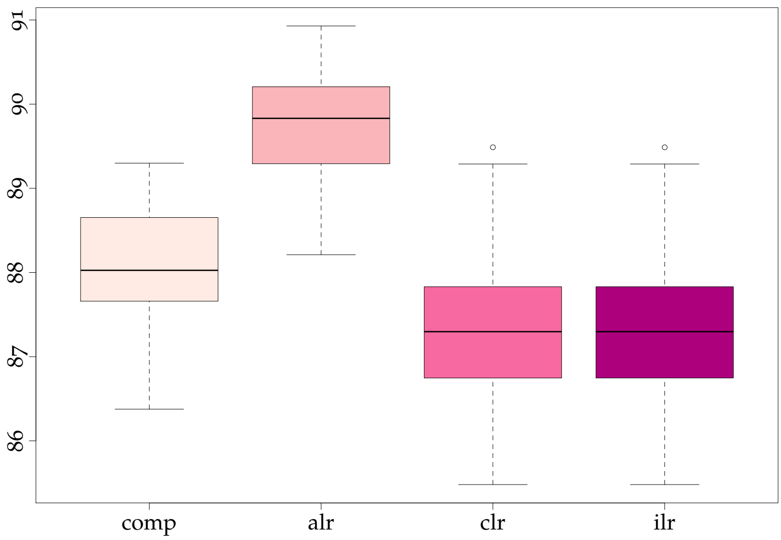

Figure 5.14: Boxplots of the percentage of correct $Ling_{[22]}$ classification with C4.5 method in Simplex space.

Similar results to the Multivariate approach, albeit slightly worse.

5.5.8 *BCCF methods*

Fuzzy space

	BCCF1	BCCF2	BCCF3	BCCF4
Minimum	89.85	89.79	89.99	89.66
Median	90.73	90.58	90.72	90.59
Mean	90.70	90.59	90.69	90.57
Maximun	91.16	91.32	91.31	91.62
Deviation	0.24	0.33	0.22	0.35

Table 5.17: Summary of the percentage of correct classification for $Ling_{[22]}$ variable with BCCF methods in Fuzzy space.

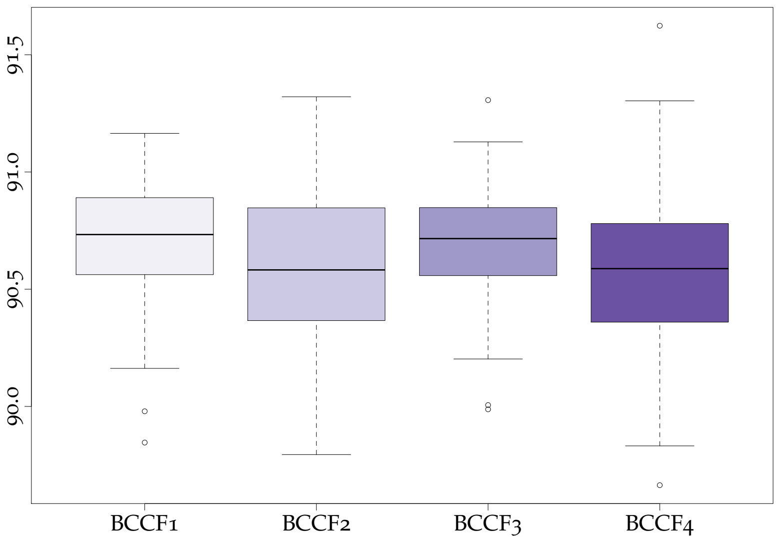

Figure 5.15: Boxplots of the percentage of correct $Ling_{[22]}$ classification with BCCF methods in Fuzzy space.

All have excellent performance, although the ones that search for δ have around 50% more variability.

5.5.9 Conclusions

For *Ling*$_{[22]}$, the best methods are LDA and LR in the Multivariate space, but only when they are presented with the 1-arity location datasets. This contrasts with similar results in BCCF1 and BCCF3 without needing to cast data for individual classifiers. The Simplex approach does not produce any noticeable improvement. The key features for inferring the linguistic characteristic are the location ones. Spreads (shape) do not introduce additional knowledge in this problem.

Table 5.18 shows how the BCCF4 method beats[14] all other algorithms in the non-Fuzzy space approach when the data is presented *as-is*.

[14] 1R is produces the same mean value but has 60% more variability.

Table 5.18: Summary of the percentage of correct classification for *Ling*$_{[22]}$ variable without transforming the data.

	BCCF4	LDA	LR	SVM	NN	*k*-NN	1R	C4.5
Minimum	89.66	89.45	88.93	87.83	86.74	87.11	89.29	86.93
Median	90.59	90.38	90.20	89.48	89.84	88.92	90.57	88.39
Mean	90.57	90.37	90.24	89.48	89.72	88.94	90.57	88.37
Maximun	91.62	91.45	91.29	90.75	91.28	90.92	91.65	89.47
Deviation	0.35	0.30	0.39	0.58	0.84	0.67	0.56	0.59

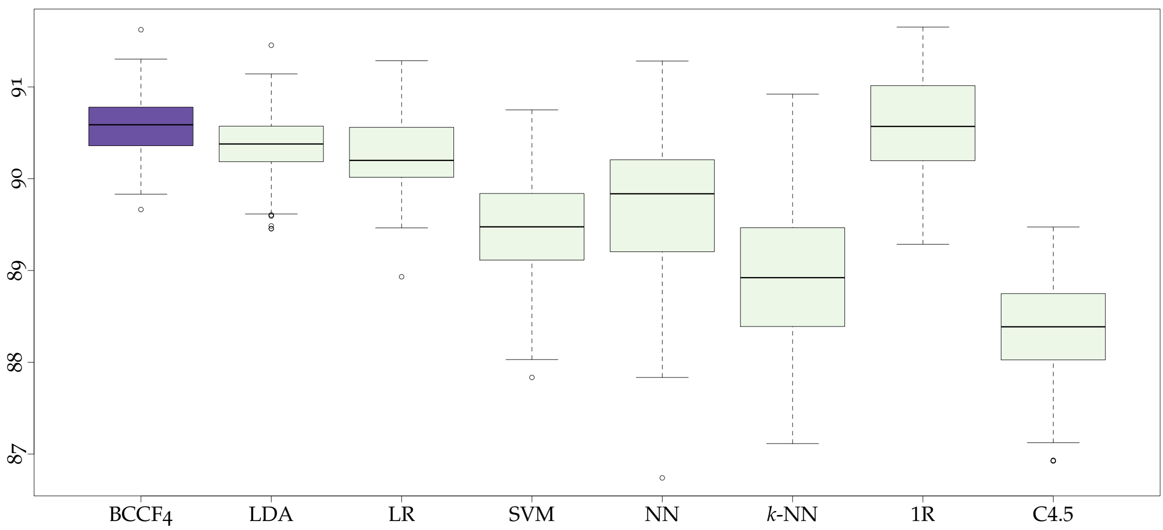

Figure 5.16: Boxplots of the percentage of correct *Ling*$_{[22]}$ classification without transforming the data.

5.6 Ling

For this problem the results are more or less the same as the subset problem $Ling_{[22]}$, although in the low 80s% scale performance wise. Much of the comments refer to the previous ones in the $Ling_{[22]}$ classification problem.

5.6.1 Linear Discriminant Analysis

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	83.69	84.28	84.28	31.78	83.57
Median	84.36	84.66	84.77	32.74	84.35
Mean	84.35	84.67	84.75	32.73	84.34
Maximun	84.89	85.09	85.22	33.89	84.87
Deviation	0.19	0.12	0.11	0.36	0.19

Table 5.19: Summary of the percentage of correct classification for *Ling* variable with LDA method in Multivariate space.

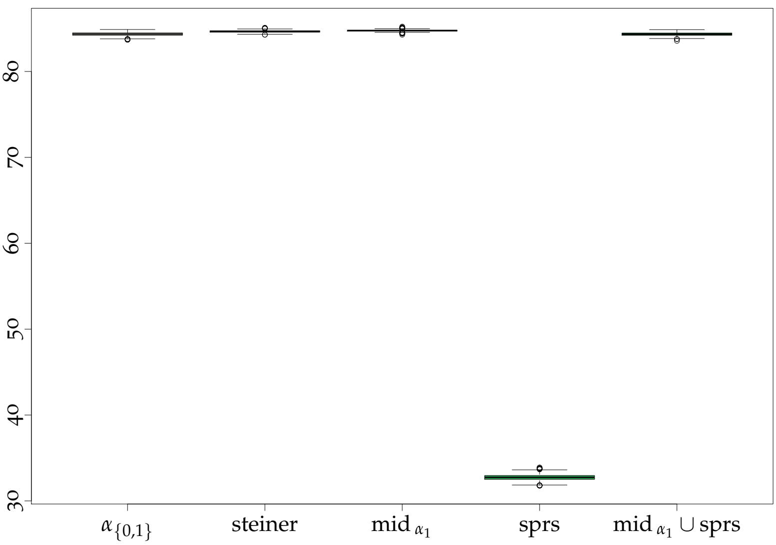

Figure 5.17: Boxplots of the percentage of correct $Ling_{[22]}$ classification with LDA method in Multivariate space.

The performance is good in all datasets but sprs, which leads to the same conclusion about location/dispersion as in the subset problem $Ling_{[22]}$. The algorithm behaves more stable in the 1-arity datasets. Notice the slight performance degradation in the mid $_{\alpha_1} \cup$ sprs with respect to mid $_{\alpha_1}$ accompanied by a relatively big increase in variability.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	83.72	62.15	62.13	62.13
Median	84.36	63.02	63.01	63.01
Mean	84.35	63.03	63.02	63.02
Maximun	84.86	63.85	63.81	63.81
Deviation	0.19	0.26	0.26	0.26

Table 5.20: Summary of the percentage of correct classification for *Ling* variable with LDA method in Simplex space.

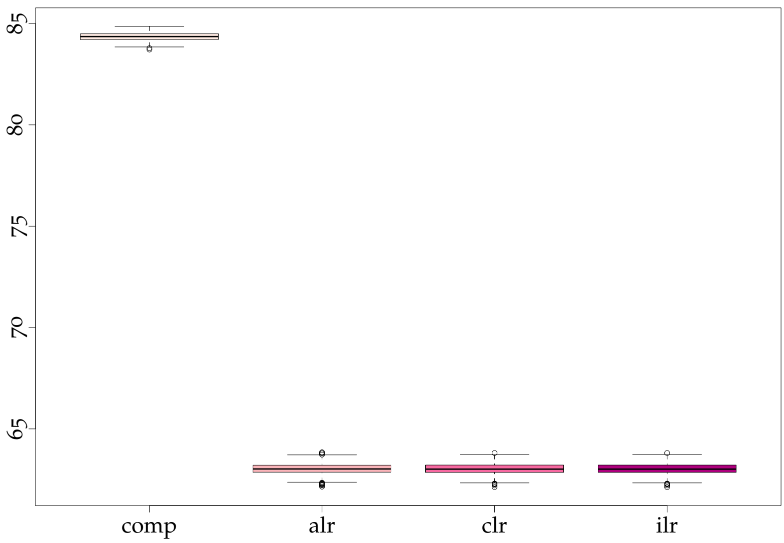

Figure 5.18: Boxplots of the percentage of correct *Ling* classification with LDA method in Simplex space.

The comp representation performs as well as in the $\alpha_{\{0,1\}}$ dataset. The bad behavior of the log-ratio transformations is the same as it was in the subset problem, cf. 5.5.1.

5.6.2 *Logistic Regression*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	84.72	84.93	85.07	31.72	84.72
Median	85.29	85.58	85.65	32.66	85.29
Mean	85.28	85.58	85.62	32.69	85.28
Maximun	85.94	86.23	85.94	33.38	85.94
Deviation	0.24	0.20	0.17	0.35	0.24

Table 5.21: Summary of the percentage of correct classification for *Ling* variable with LR method in Multivariate space.

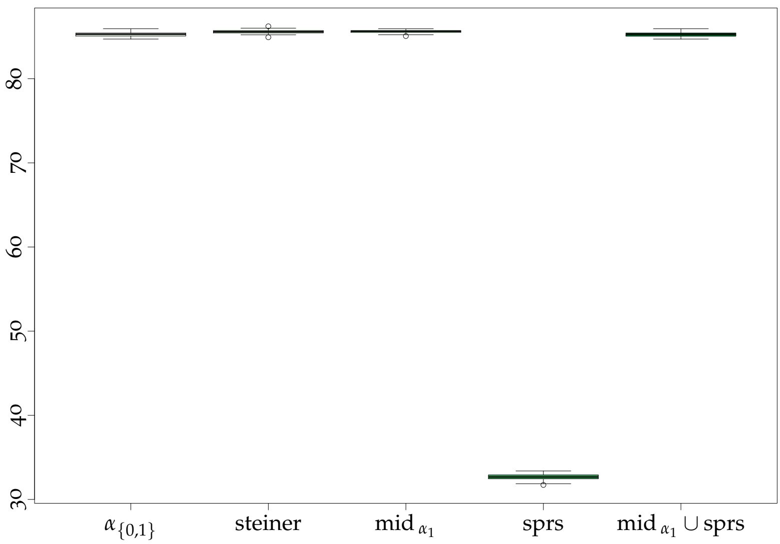

Figure 5.19: Boxplots of the percentage of correct *Ling* classification with LR method in Multivariate space.

Similar behaviors as LDA. The degradation in variability is only half of what we had for the subset problem $Ling_{[22]}$.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	84.72	84.57	84.57	84.57
Median	85.29	85.22	85.22	85.22
Mean	85.28	85.21	85.21	85.21
Maximun	85.94	85.73	85.73	85.73
Deviation	0.24	0.23	0.23	0.23

Table 5.22: Summary of the percentage of correct classification for *Ling* variable with LR method in Simplex space.

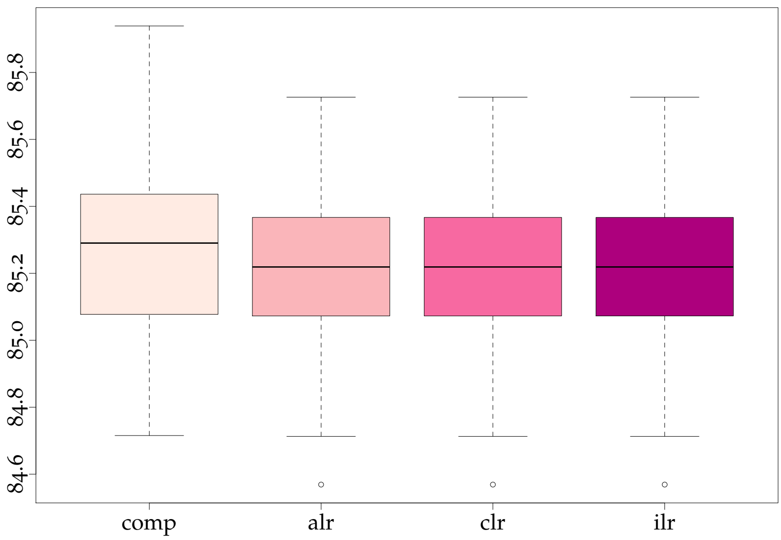

Figure 5.20: Boxplots of the percentage of correct *Ling* classification with LR method in Simplex space.

Marginally worse performance than with the Multivariate approach.

5.6.3 *Support Vector Machines*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	80.03	84.57	85.00	32.23	81.90
Median	81.18	85.22	85.51	33.45	82.70
Mean	81.10	85.19	85.52	33.53	82.67
Maximun	81.90	85.66	86.01	34.54	83.56
Deviation	0.39	0.27	0.22	0.52	0.36

Table 5.23: Summary of the percentage of correct classification for *Ling* variable with SVM method in Multivariate space.

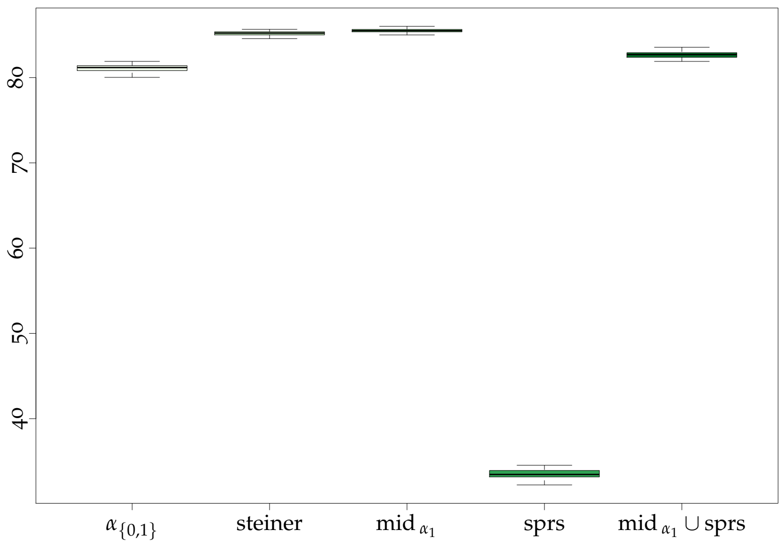

Figure 5.21: Boxplots of the percentage of correct *Ling* classification with SVM method in Multivariate space.

Same comments as the subset case cf. 5.5.3.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	84.50	83.63	83.56	83.56
Median	84.93	84.14	84.39	84.39
Mean	84.96	84.19	84.37	84.37
Maximun	85.36	84.86	85.00	85.00
Deviation	0.18	0.28	0.29	0.29

Table 5.24: Summary of the percentage of correct classification for *Ling* variable with SVM method in Simplex space.

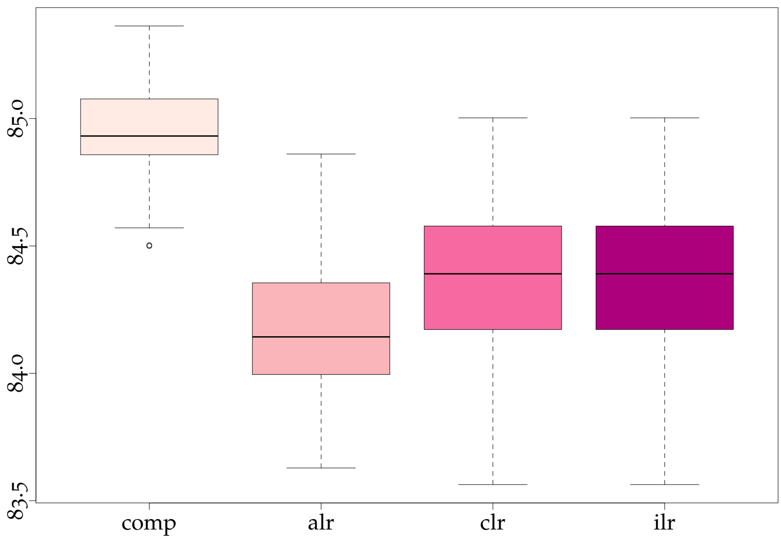

Figure 5.22: Boxplots of the percentage of correct *Ling* classification with SVM method in Simplex space.

Same behavior as the one seen in the subset problem in 5.5.3. An increase in stability in the comp representation with respect to the $\alpha_{\{0,1\}}$.

5.6.4 *Neural Network*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}\cup$ sprs
Minimum	83.19	73.26	73.12	29.78	82.84
Median	84.75	78.15	77.83	31.29	84.35
Mean	84.69	78.01	77.91	31.34	84.30
Maximun	85.80	82.69	82.90	33.74	85.37
Deviation	0.57	1.86	1.87	0.74	0.47

Table 5.25: Summary of the percentage of correct classification for *Ling* variable with NN method in Multivariate space.

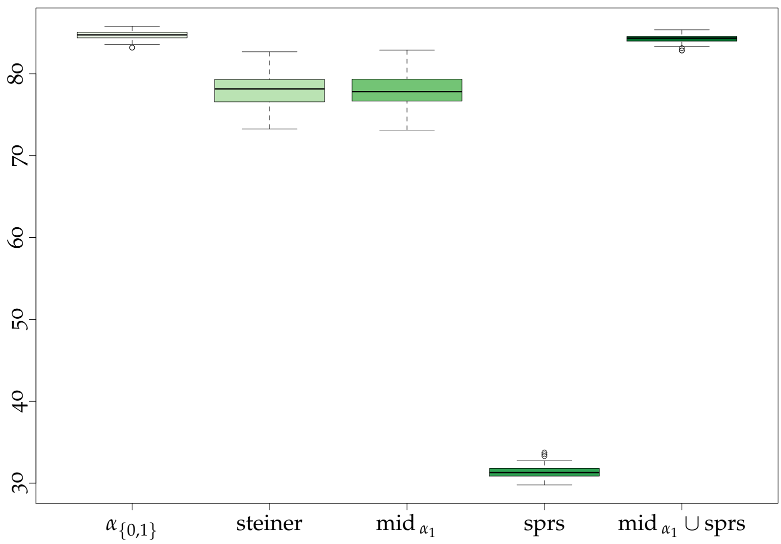

Figure 5.23: Boxplots of the percentage of correct *Ling* classification with NN method in Multivariate space.

Same comment in the 85% performance scale as 5.5.4.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	82.70	79.38	79.67	79.67
Median	84.28	82.59	82.69	82.69
Mean	84.25	82.57	82.56	82.56
Maximun	85.73	85.15	84.93	84.93
Deviation	0.56	1.06	1.01	1.01

Table 5.26: Summary of the percentage of correct classification for *Ling* variable with NN method in Simplex space.

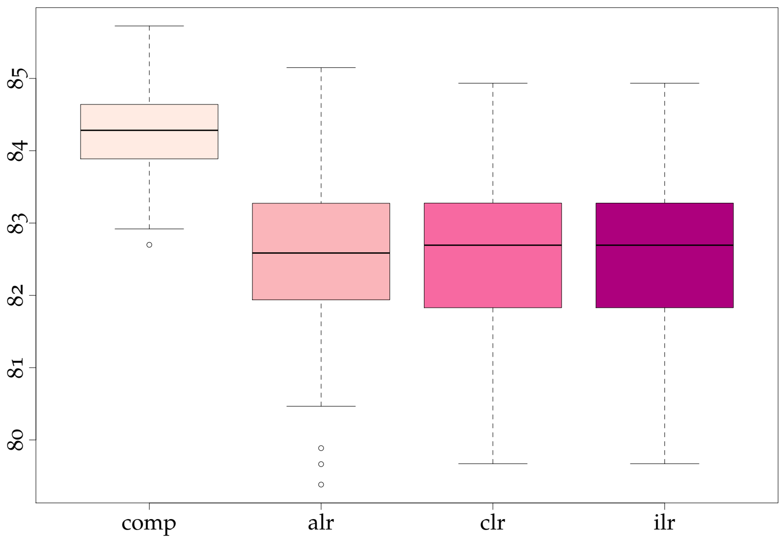

Figure 5.24: Boxplots of the percentage of correct *Ling* classification with NN method in Simplex space.

Same relative behavior as in 5.5.4 Simplex space stanza.

5.6.5 *k-Nearest Neighbor*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}$ ∪ sprs
Minimum	82.11	82.77	81.76	29.28	81.61
Median	83.56	83.63	82.80	31.22	82.70
Mean	83.53	83.63	82.80	31.24	82.70
Maximun	84.43	84.49	83.71	32.95	84.06
Deviation	0.41	0.37	0.42	0.69	0.41

Table 5.27: Summary of the percentage of correct classification for *Ling* variable with *k*-NN method in Multivariate space.

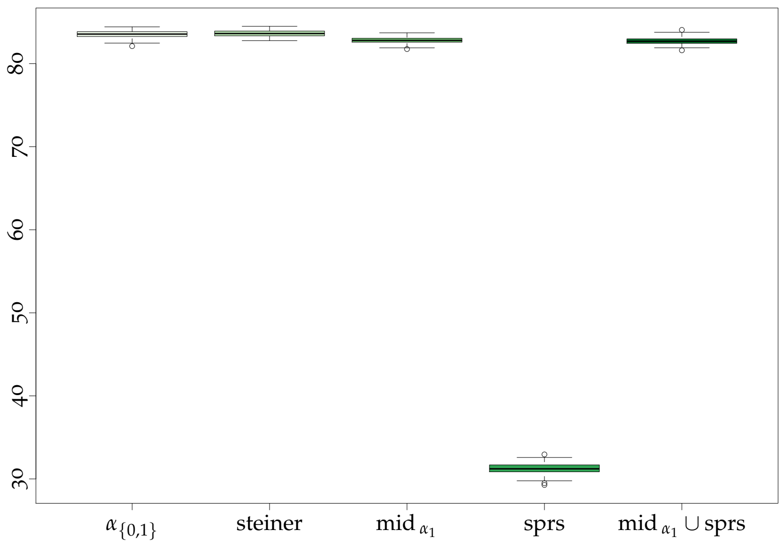

Figure 5.25: Boxplots of the percentage of correct *Ling* classification with *k*-NN method in Multivariate space.

We have good results in the Multivariate space. In contrast with the subset problem 5.5.5, there is no degradation in the redundant dataset mid $_{\alpha_1}$ ∪ sprs with respect to mid $_{\alpha_1}$. This leads to the conclusion that the influence of the noise component is less relevant for this problem when we have more instances.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	81.62	81.32	80.89	80.89
Median	82.55	82.34	81.62	81.62
Mean	82.55	82.40	81.70	81.70
Maximun	83.56	83.78	82.84	82.84
Deviation	0.42	0.53	0.43	0.43

Table 5.28: Summary of the percentage of correct classification for *Ling* variable with *k*-NN method in Simplex space.

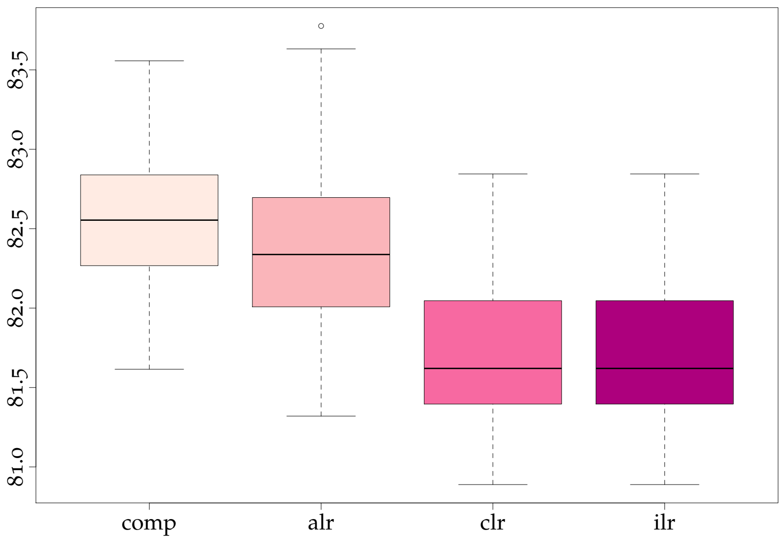

Figure 5.26: Boxplots of the percentage of correct *Ling* classification with *k*-NN method in Simplex space.

Same as 5.5.5 Simplex stanza, a bit worse than simpler data configurations.

5.6.6 1 – Rules

– Multivariate space

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	82.91	83.20	83.05	27.04	83.05
Median	84.28	84.36	84.00	29.31	84.00
Mean	84.21	84.33	84.01	29.25	84.01
Maximun	85.44	85.44	84.92	31.87	84.92
Deviation	0.55	0.41	0.37	1.03	0.37

Table 5.29: Summary of the percentage of correct classification for *Ling* variable with 1R method in Multivariate space.

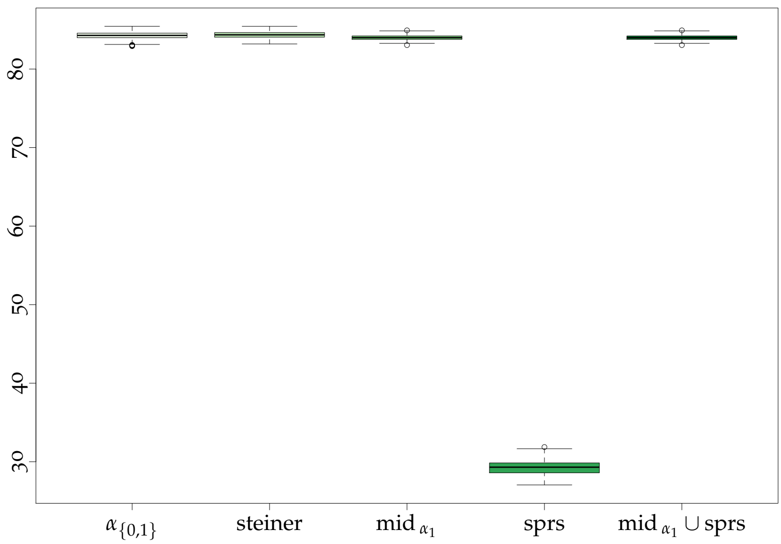

Figure 5.27: Boxplots of the percentage of correct *Ling* classification with 1R method in Multivariate space.

Good results on par with more complicated methods, same comment as 5.5.6.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	81.39	83.71	64.31	64.31
Median	82.73	84.86	66.47	66.47
Mean	82.72	84.84	66.40	66.40
Maximun	83.79	85.51	68.06	68.06
Deviation	0.48	0.34	0.79	0.79

Table 5.30: Summary of the percentage of correct classification for *Ling* variable with 1R method in Simplex space.

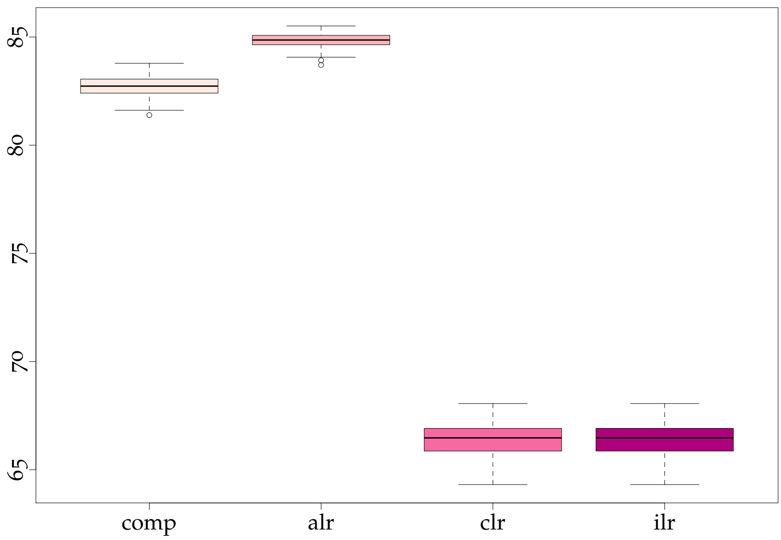

Figure 5.28: Boxplots of the percentage of correct *Ling* classification with 1R method in Simplex space.

Same issues in the clr and ilr log-ratio transformations that were encountered in the subset problem, see 5.5.6.

5.6.7 C4.5

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}$ ∪ sprs
Minimum	82.84	84.42	83.77	29.42	83.35
Median	83.70	85.15	84.72	31.22	84.57
Mean	83.71	85.16	84.71	31.21	84.54
Maximun	84.50	85.80	85.58	33.60	85.51
Deviation	0.37	0.34	0.38	0.96	0.46

Table 5.31: Summary of the percentage of correct classification for *Ling* variable with C4.5 method in Multivariate space.

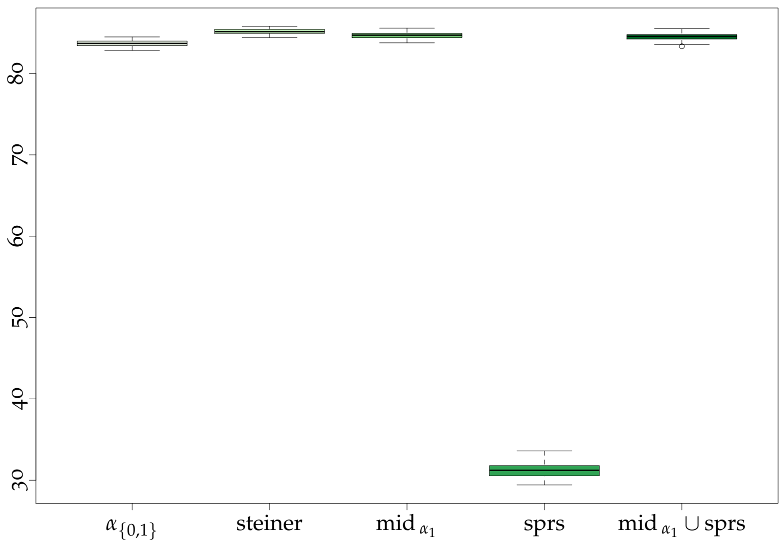

Figure 5.29: Boxplots of the percentage of correct *Ling* classification with C4.5 method in Multivariate space.

Similar results as in the simpler 1R. Same problem as the one described in 5.5.7, although in this case the C4.5 is much more stable in $\alpha_{\{0,1\}}$ representation than in the 1R classifier.

- **Simplex space**

	comp	alr	clr	ilr
Minimum	80.97	84.43	79.23	79.23
Median	82.77	85.87	81.40	81.40
Mean	82.76	85.87	81.38	81.38
Maximun	83.77	86.45	82.55	82.55
Deviation	0.53	0.36	0.59	0.59

Table 5.32: Summary of the percentage of correct classification for *Ling* variable with C4.5 method in Simplex space.

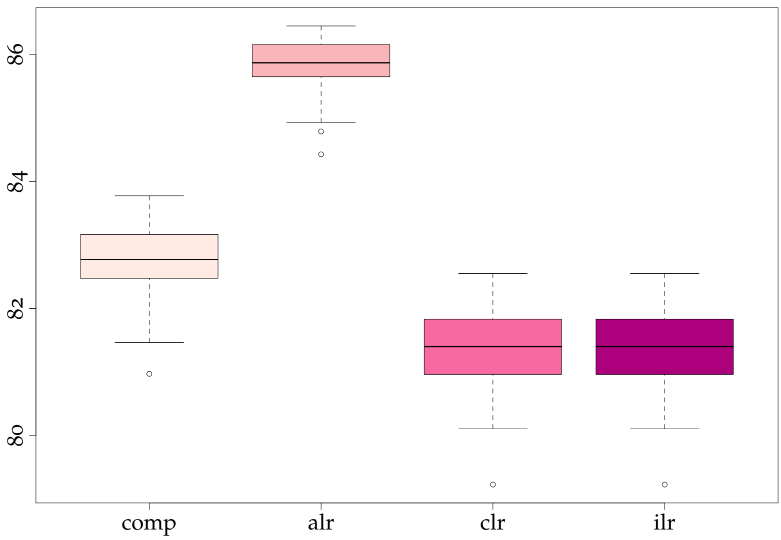

Figure 5.30: Boxplots of the percentage of correct *Ling* classification with C4.5 method in Simplex space.

Same behavior as in 5.5.7 Simplex stanza.

5.6.8 BCCF methods

	BCCF1	BCCF2	BCCF3	BCCF4
Minimum	85.53	84.94	85.50	84.92
Median	85.88	85.59	85.88	85.62
Mean	85.88	85.60	85.89	85.60
Maximun	86.39	86.30	86.31	86.03
Deviation	0.17	0.25	0.18	0.21

Table 5.33: Summary of the percentage of correct classification for *Ling* variable with BCCF methods in Fuzzy space.

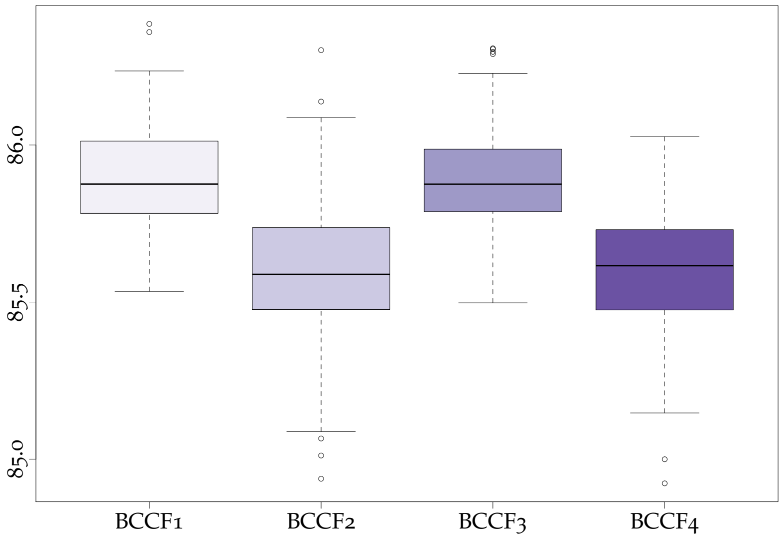

Figure 5.31: Boxplots of the percentage of correct *Ling* classification with BCCF methods in Fuzzy space.

Good results. The method achieves the best result for the *Ling* problem. Again, the ones that not look for δ have the best behavior, cf. 5.5.8.

5.6.9 Conclusions

The issues described in 5.5.9 also apply here. The BCCF4 method is the best when data is not transformed for the classifier.

Table 5.34: Summary of the percentage of correct classification for *Ling* variable without transforming the data.

	BCCF4	LDA	LR	SVM	NN	*k*-NN	1R	C4.5
Minimum	84.92	83.69	84.72	80.03	83.19	82.11	82.91	82.84
Median	85.62	84.36	85.29	81.18	84.75	83.56	84.28	83.70
Mean	85.60	84.35	85.28	81.10	84.69	83.53	84.21	83.71
Maximun	86.03	84.89	85.94	81.90	85.80	84.43	85.44	84.50
Deviation	0.21	0.19	0.24	0.39	0.57	0.41	0.55	0.37

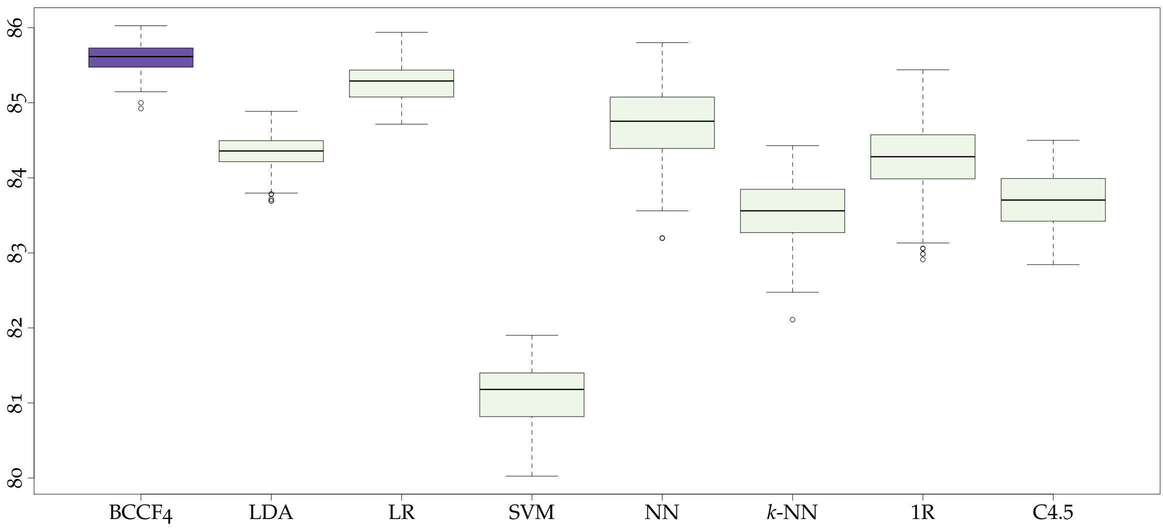

Figure 5.32: Boxplots of the percentage of correct *Ling* classification without transforming the data.

5.7 Sex *classification problem*

5.7.1 *Linear Discriminant Analysis*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	67.88	45.34	45.27	66.48	67.88
Median	69.39	50.53	50.21	68.98	68.92
Mean	69.32	50.42	49.70	68.86	69.10
Maximun	70.32	52.55	51.63	71.23	71.32
Deviation	0.47	1.52	1.51	1.06	0.78

Table 5.35: Summary of the percentage of correct classification for *Sex* variable with LDA method in Multivariate space.

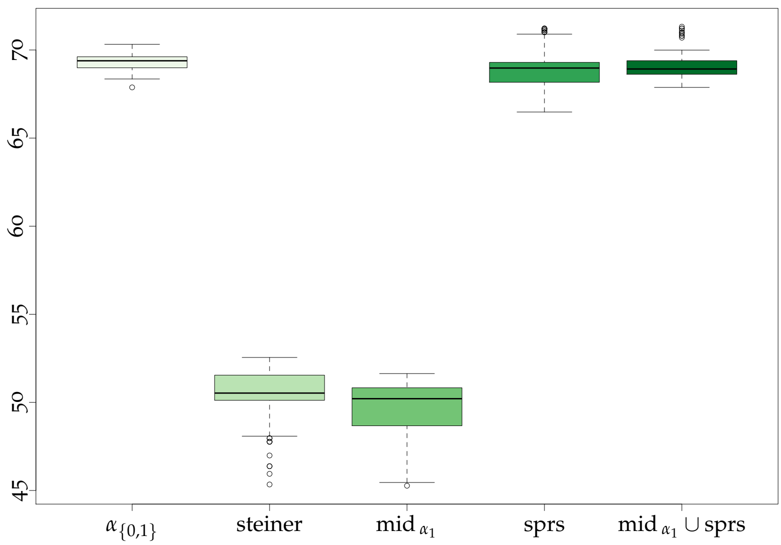

Figure 5.33: Boxplots of the percentage of correct *Sex* classification with LDA method in Multivariate space.

Moderately good results in the datasets on which an amplitude measure could be deduced. No inference possible in the 1-arity location scenario[15].

[15] In this problem the situation reverses with respect to the linguistic problem. Location variables individually offer no information, and the 1-arity central datasets steiner, mid $_{\alpha_1}$ will be useless for trying to explain the *Sex* variable.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	66.78	71.11	71.42	71.42
Median	68.90	72.83	72.69	72.69
Mean	68.87	72.80	72.71	72.71
Maximun	70.82	75.06	73.84	73.84
Deviation	0.89	1.00	0.56	0.56

Table 5.36: Summary of the percentage of correct classification for *Sex* variable with LDA method in Simplex space.

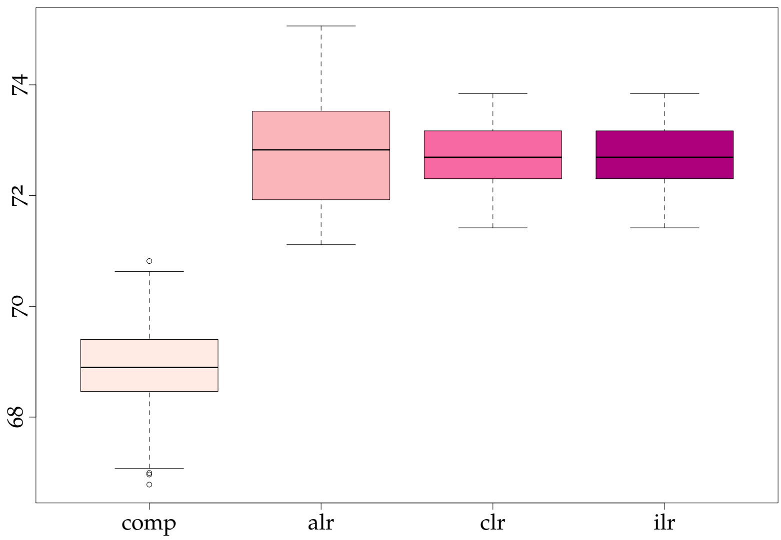

Figure 5.34: Boxplots of the percentage of correct *Sex* classification with LDA method in Simplex space.

Improvement in the log-ratio transformations with respect to the Multivariate approach, the clr , ilr offer much more stability than the alr transformation.

5.7.2 *Logistic Regression*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}\cup$ sprs
Minimum	67.47	44.65	44.65	67.68	67.47
Median	69.80	50.81	50.71	69.80	69.80
Mean	69.53	50.22	50.02	69.59	69.53
Maximun	70.71	51.92	51.82	70.71	70.71
Deviation	0.72	1.44	1.53	0.75	0.72

Table 5.37: Summary of the percentage of correct classification for *Sex* variable with LR method in Multivariate space.

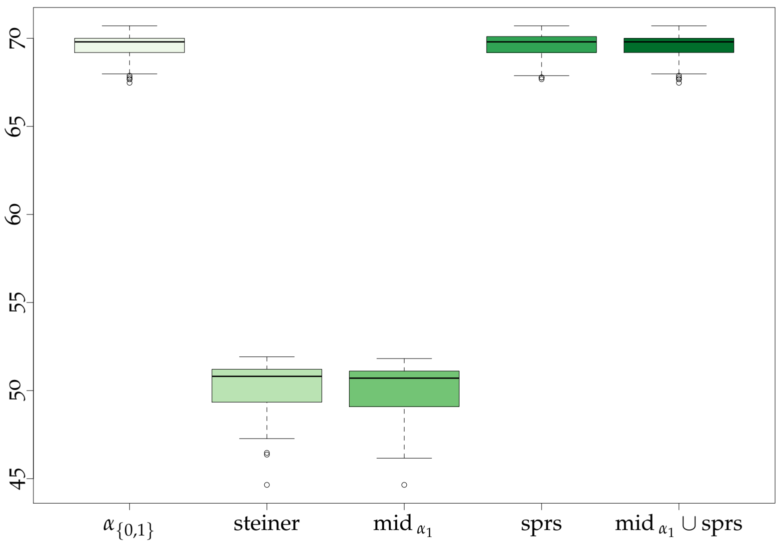

Figure 5.35: Boxplots of the percentage of correct *Sex* classification with LR method in Multivariate space.

Similar mean values results as the LDA Multivariate case, but less variability in the sprs datasets, situation that it's the opposite for the $\alpha_{\{0,1\}}$ dataset.

- **Simplex space**

	comp	alr	clr	ilr
Minimum	67.47	71.72	71.72	71.72
Median	69.80	73.64	73.64	73.64
Mean	69.53	73.69	73.69	73.69
Maximun	70.71	75.35	75.35	75.35
Deviation	0.72	0.91	0.91	0.91

Table 5.38: Summary of the percentage of correct classification for *Sex* variable with LR method in Simplex space.

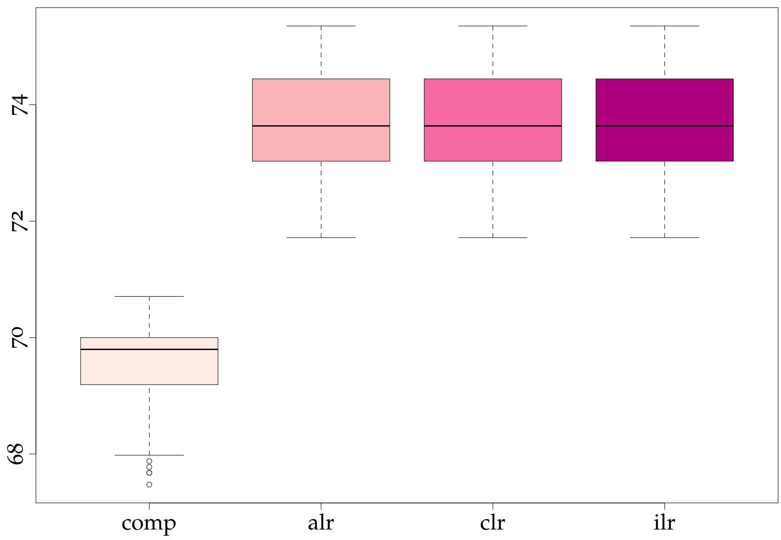

Figure 5.36: Boxplots of the percentage of correct *Sex* classification with LR method in Simplex space.

Improvement in the log-ratio transformation with more instability with respect to the Multivariate case.

5.7.3 *Support Vector Machines*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	65.56	45.25	43.54	72.83	66.36
Median	67.88	48.74	49.44	75.56	68.03
Mean	67.86	48.96	49.50	75.64	68.11
Maximun	70.51	54.34	54.04	77.68	71.11
Deviation	1.08	2.09	2.28	0.99	1.03

Table 5.39: Summary of the percentage of correct classification for *Sex* variable with SVM method in Multivariate space.

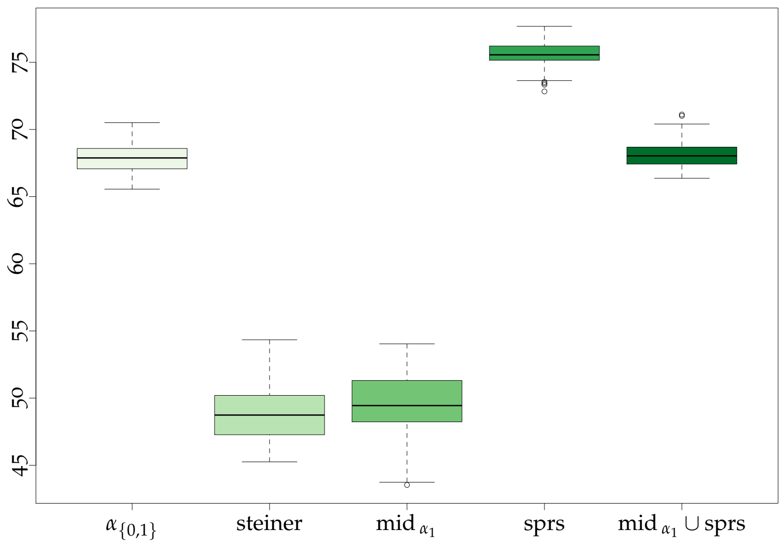

Figure 5.37: Boxplots of the percentage of correct *Sex* classification with SVM method in Multivariate space.

Good results with the sprs. A noticeable degradation with its superset. The method does not fullylearn the concept of spread from the $\alpha_{\{0,1\}}$ dataset.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	56.06	71.82	73.13	73.13
Median	58.08	74.24	74.95	74.95
Mean	58.05	74.12	74.96	74.96
Maximun	59.70	76.26	77.17	77.17
Deviation	0.79	1.04	0.98	0.98

Table 5.40: Summary of the percentage of correct classification for *Sex* variable with SVM method in Simplex space.

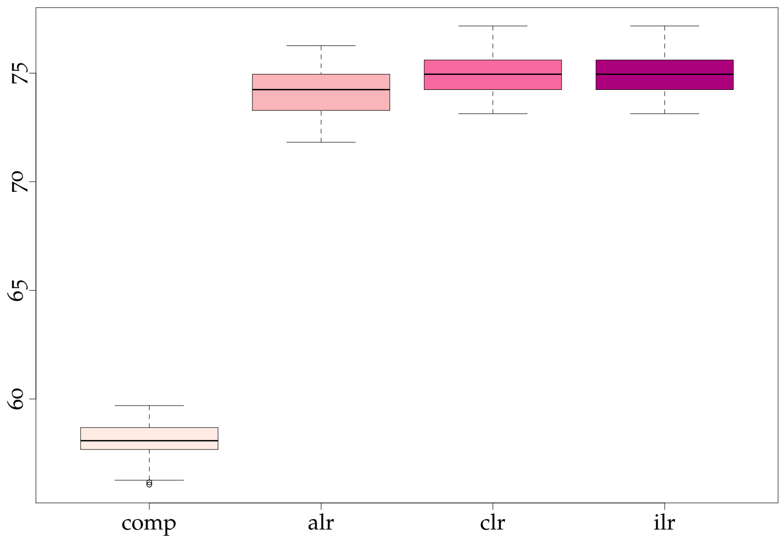

Figure 5.38: Boxplots of the percentage of correct *Sex* classification with SVM method in Simplex space.

Good results for the log-ratio transformations, bad results for the comp dataset.

5.7.4 *Neural Network*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}$ $\cup$ sprs
Minimum	67.58	48.79	48.69	72.02	70.51
Median	70.81	50.00	50.10	74.04	73.23
Mean	70.96	50.02	50.07	74.07	73.22
Maximun	74.04	51.52	51.41	75.96	76.26
Deviation	1.22	0.55	0.52	0.86	1.02

Table 5.41: Summary of the percentage of correct classification for *Sex* variable with NN method in Multivariate space.

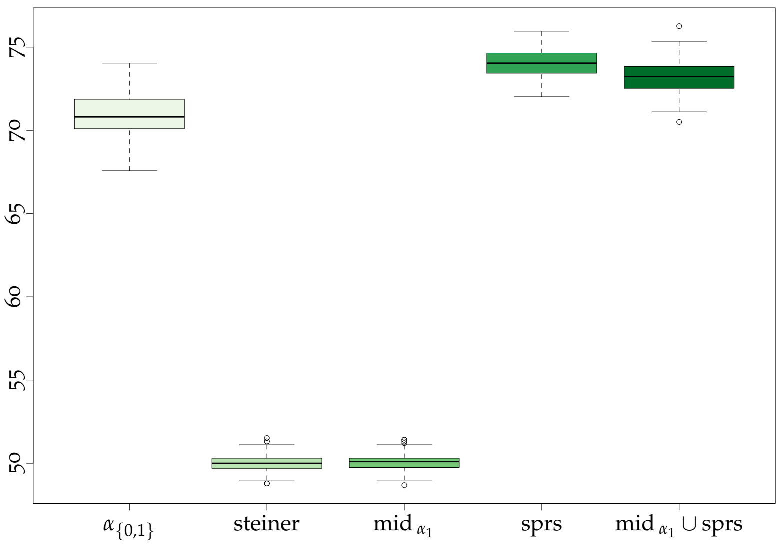

Figure 5.39: Boxplots of the percentage of correct *Sex* classification with NN method in Multivariate space.

As with the SVM, the sprs information behaves the best, the superset degrades, and it's not able to form internally the full concept of spread from the α-cuts representation.

- **Simplex space**

Table 5.42: Summary of the percentage of correct classification for *Sex* variable with NN method in Simplex space.

	comp	alr	clr	ilr
Minimum	71.11	71.82	70.40	70.40
Median	73.43	74.14	73.94	73.94
Mean	73.42	74.17	73.86	73.86
Maximun	75.96	76.97	76.16	76.16
Deviation	1.07	1.06	1.14	1.14

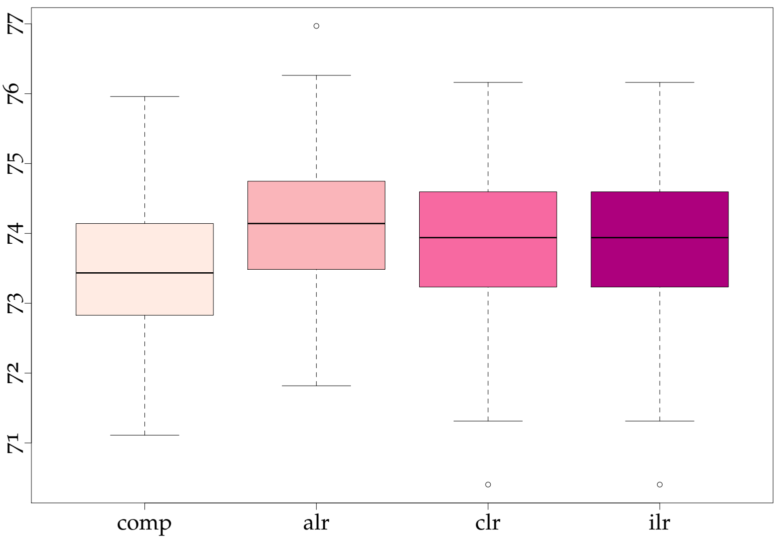

Figure 5.40: Boxplots of the percentage of correct *Sex* classification with NN method in Simplex space.

Here we have the opposite behavior in the SVM with respect to the divergence in the comp, $\alpha_{\{0,1\}}$ representations. And an improvement over the Multivariate representation, although the effect is not as severe as with the SVM algorithm.

5.7.5 *k-Nearest Neighbor*

– **Multivariate space**

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	62.02	44.24	45.66	70.40	70.61
Median	65.96	47.68	49.39	72.93	73.38
Mean	65.97	47.89	49.37	73.02	73.23
Maximun	69.80	51.92	53.74	75.66	75.35
Deviation	1.64	1.75	1.59	1.03	1.07

Table 5.43: Summary of the percentage of correct classification for *Sex* variable with *k*-NN method in Multivariate space.

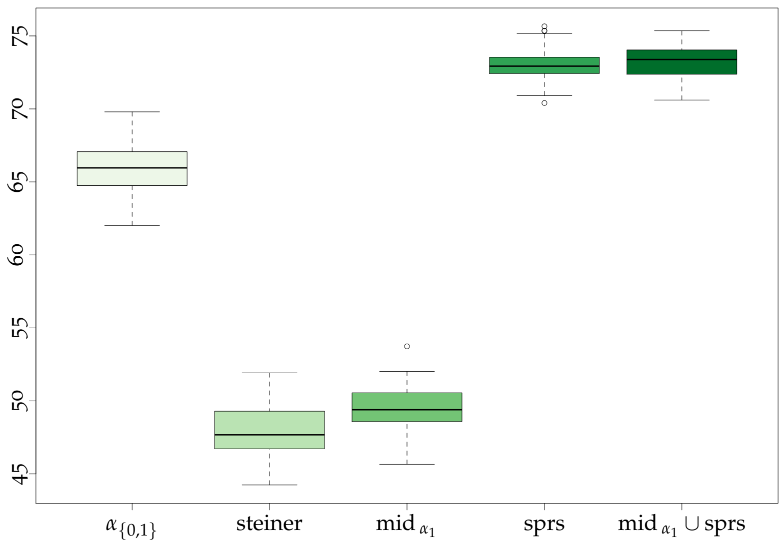

Figure 5.41: Boxplots of the percentage of correct *Sex* classification with *k*-NN method in Multivariate space.

The sprs and it's superset are the ones that produce the best results for this representation and algorithm, the influence of the spurious central data is not enough to degrade the method.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	69.90	70.10	71.41	71.41
Median	72.98	72.83	74.09	74.09
Mean	72.75	72.87	74.13	74.13
Maximun	76.06	74.85	76.57	76.57
Deviation	1.25	0.99	1.07	1.07

Table 5.44: Summary of the percentage of correct classification for *Sex* variable with *k*-NN method in Simplex space.

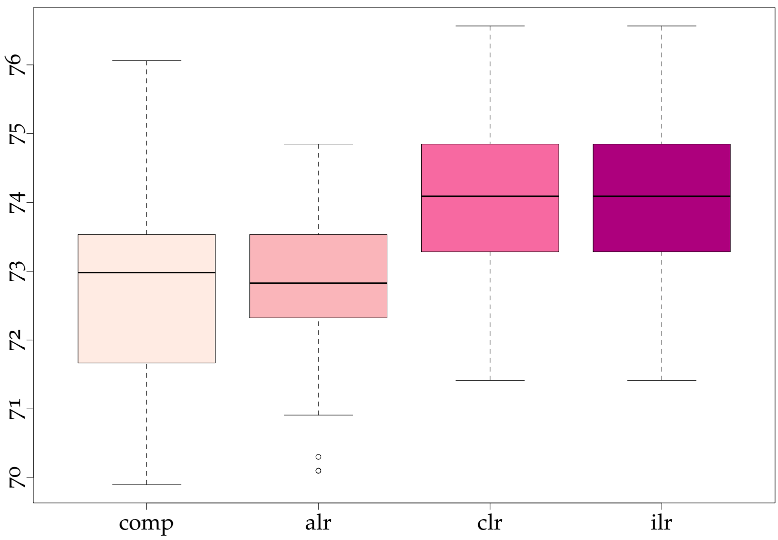

Figure 5.42: Boxplots of the percentage of correct *Sex* classification with *k*-NN method in Simplex space.

Improvement in the comp dataset over the $\alpha_{\{0,1\}}$ in mean value and stability, good results for the log-ratio transformations.

5.7.6 1 – Rules

– Multivariate space

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1}$ $\cup$ sprs
Minimum	49.80	44.75	44.75	68.18	68.18
Median	52.93	48.94	50.10	70.71	70.71
Mean	52.93	49.10	49.88	70.64	70.64
Maximun	56.77	52.93	54.55	72.73	72.73
Deviation	1.56	1.97	1.91	1.03	1.03

Table 5.45: Summary of the percentage of correct classification for *Sex* variable with 1R method in Multivariate space.

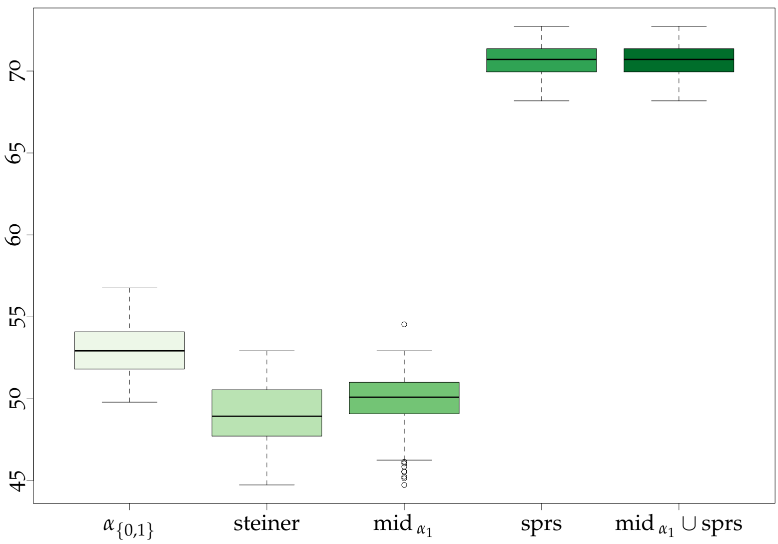

Figure 5.43: Boxplots of the percentage of correct *Sex* classification with 1R method in Multivariate space.

Due to the nature of the method it cannot build the concept of spread from $\alpha_{\{0,1\}}$ so it fails for that representation. With the sprs representation, it produces acceptable results, which again due to it's nature do not degrade in the noisy superset.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	68.28	56.57	66.16	66.16
Median	71.01	60.91	70.20	70.20
Mean	71.03	60.84	70.10	70.10
Maximun	73.13	64.44	73.13	73.13
Deviation	1.01	1.75	1.45	1.45

Table 5.46: Summary of the percentage of correct classification for *Sex* variable with 1R method in Simplex space.

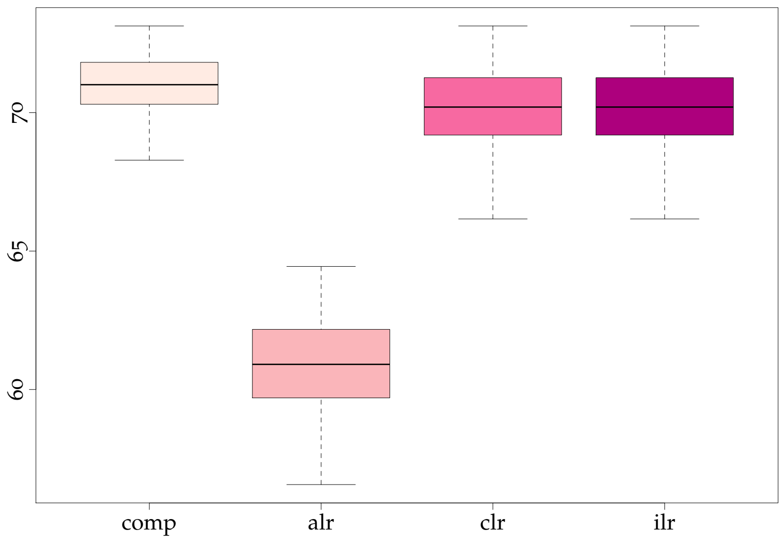

Figure 5.44: Boxplots of the percentage of correct *Sex* classification with 1R method in Simplex space.

In the comp representation, it succeeds in contrast with the $\alpha_{\{0,1\}}$ representation, due to the fact that the comp representation barring the first variable $\inf 0$, all coordinates contain relative difference values—i.e. the difference with the previous datum[16]. No improvement in the log-ratio transformations, with noticeable degradation in the alr.

[16] Which also happens to $\inf 0$, which in reality is $\inf 0 - 0$ in the comp representation.

5.7.7 C4.5

– Multivariate space

	$\alpha_{\{0,1\}}$	steiner	mid $_{\alpha_1}$	sprs	mid $_{\alpha_1} \cup$ sprs
Minimum	49.09	49.49	49.49	71.21	71.11
Median	50.10	49.49	49.49	73.84	73.64
Mean	50.52	49.49	49.49	73.83	73.70
Maximun	53.33	49.49	49.49	76.36	76.36
Deviation	1.20	0.00	0.00	0.99	1.06

Table 5.47: Summary of the percentage of correct classification for *Sex* variable with C4.5 method in Multivariate space.

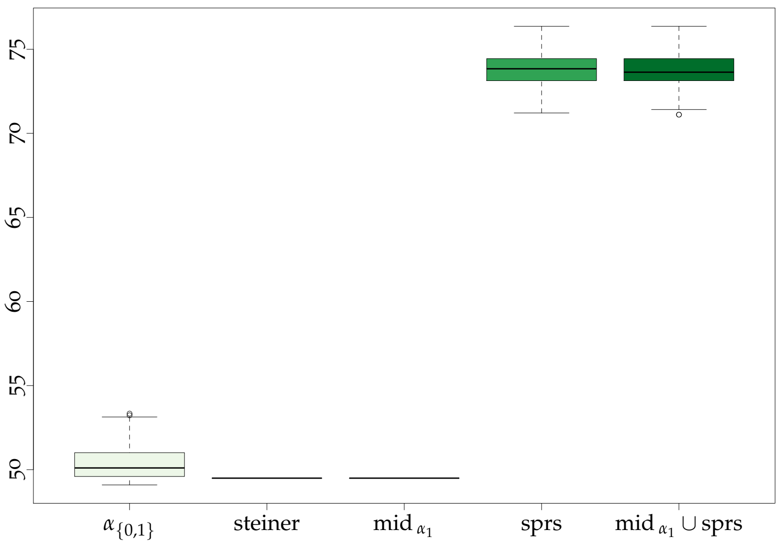

Figure 5.45: Boxplots of the percentage of correct *Sex* classification with C4.5 method in Multivariate space.

Textbook example of the shortcomings of the C4.5, it does not know how to perform addition, so it cannot build the concept of spread for the $\alpha_{\{0,1\}}$ representation. Good result with the sprs dataset, with a slight degradation in its superset.

– **Simplex space**

	comp	alr	clr	ilr
Minimum	70.61	66.77	70.61	70.61
Median	73.94	70.15	72.73	72.73
Mean	73.87	70.35	73.00	73.00
Maximun	76.16	73.64	76.77	76.77
Deviation	1.08	1.37	1.39	1.39

Table 5.48: Summary of the percentage of correct classification for *Sex* variable with C4.5 method in Simplex space.

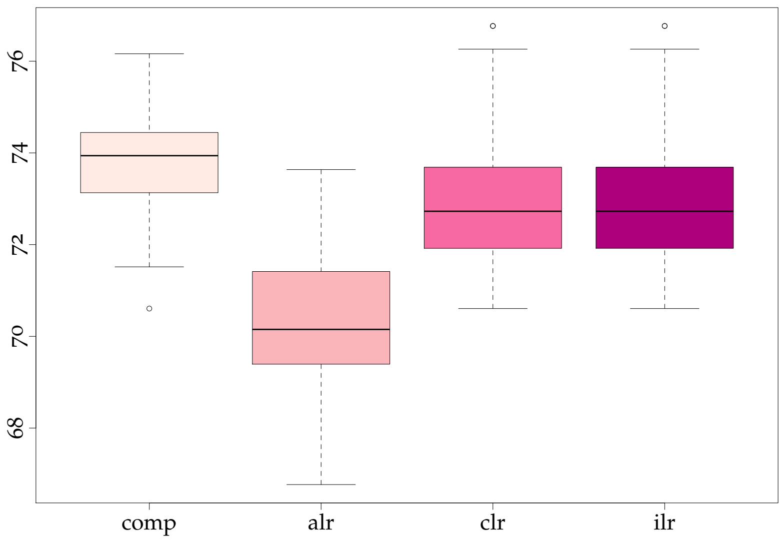

Figure 5.46: Boxplots of the percentage of correct *Sex* classification with C4.5 method in Simplex space.

In contraposition with the Multivariate case, the comp dataset—which is formed with ’spreads’—provides the information that the tree needs to build the classifier. A tad worse results and variability in the log ratio transformations.

5.7.8 BCCF methods

	BCCF1	BCCF2	BCCF3	BCCF4
Minimum	45.55	53.41	61.12	67.69
Median	50.45	57.89	62.28	70.42
Mean	50.30	57.68	62.27	70.32
Maximun	53.41	60.54	63.64	72.32
Deviation	1.45	1.49	0.60	0.96

Table 5.49: Summary of the percentage of correct classification for *Sex* variable with BCCF methods in Fuzzy space.

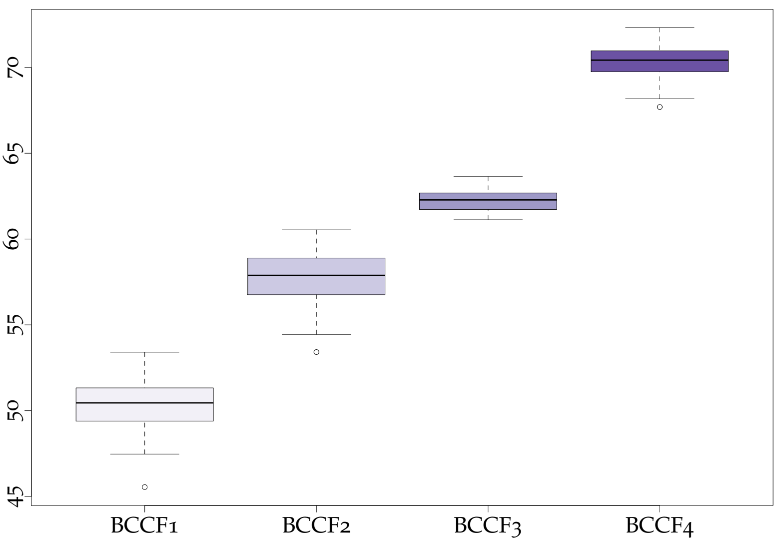

Figure 5.47: Boxplots of the percentage of correct *Sex* classification with BCCF methods in Fuzzy space.

BCCF1 and to a lesser extent BCCF2 do not work, due to the weight in the distance component of the difference between mid levels. In BCCF3, it can be minimized the influence of the the mid component in the distance and there is a slight improvement. With BCCF4, the performance is good in the lower band, 1R beats it when using the sprs representation. Although this argument could be reversed and said that the classical algorithms only beat BCCF4 when the correct information is presented. Because it beats all algorithms—but the neural network—that use the raw data $\alpha_{\{0,1\}}$ in this problem.

cf. Section 2.2

5.7.9 *Conclusions*

Of all of the algorithms with the data *as-is* only the NN method can compete in results.

Table 5.50: Summary of the percentage of correct classification for *Sex* variable without transforming the data.

	BCCF4	LDA	LR	SVM	NN	*k*-NN	1R	C4.5
Minimum	67.69	67.88	67.47	65.56	67.58	62.02	49.80	49.80
Median	70.42	69.39	69.80	67.88	70.81	65.96	52.93	52.93
Mean	70.32	69.32	69.53	67.86	70.96	65.97	52.93	52.93
Maximun	72.32	70.32	70.71	70.51	74.04	69.80	56.77	56.77
Deviation	0.96	0.47	0.72	1.08	1.22	1.64	1.56	1.56

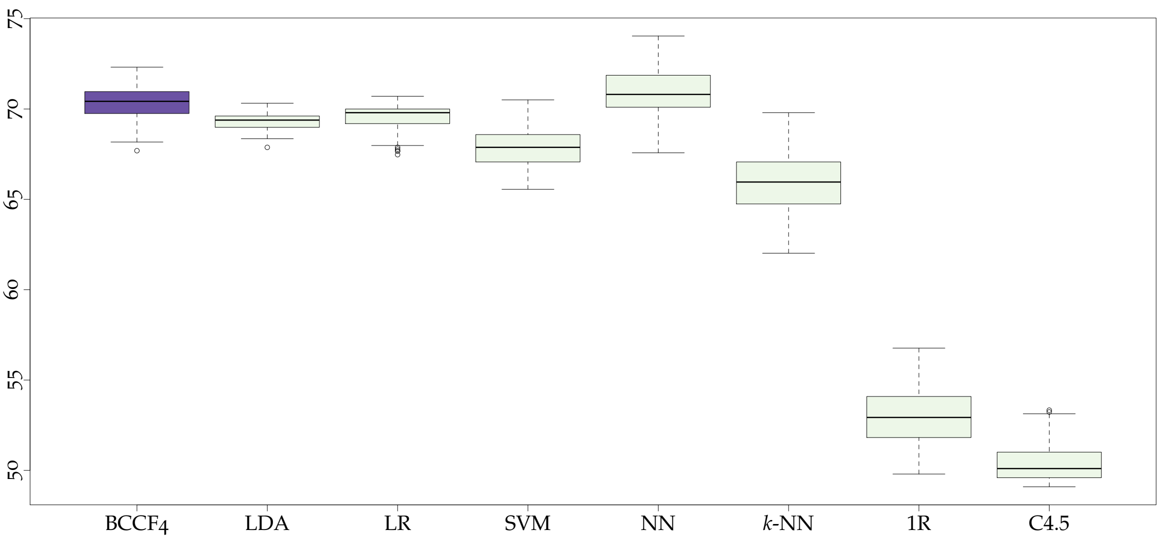

Figure 5.48: Boxplots of the percentage of correct *Sex* classification without transforming the data.

Contrast C4.5 (same argument could be said for 1R) which is useless for this problem and representation, when we spoon feed the algorithm the sprs representation, cf. 5.7.7.

The key features for inferring (albeit with less success than the linguistic problem) the genre characteristic are the linear combinations of the features that create dispersion measures. Location do not introduce additional knowledge in this problem.

5.8 Combined classification results

In this seccion we combine the means—with the geometric mean—results of the *Ling* and *Sex* problem to give an overall picture of representation, algorithm pairs performance.

The log-ratio transformations shine here, producing the overall best result with SVM followed by LR. The BCCF methods hold up, considering the data *as-is*, the BCCF4 is the best under this combination criteria closely followed by NN.

One thing to notice about the combined results is that they are a bit misleading as no deviation measure is provided, the methods based on information gaining around a feature, i.e. 1R, C4.5, are useless when trying to learn based on the concept of dispersion. This is somewhat hidden when combined with a classification result around the opposite concept, location, for which they prove very much adequate.

Table 5.51: *Ling Sex* geometric mean classification in Multivariate and Simplex space.

	LDA	LR	SVM	NN	k-NN	1R	C4.5
$\alpha_{\{0,1\}}$	76.47	77.00	74.18	77.52	74.23	66.76	65.03
steiner	65.34	65.56	64.58	62.47	63.29	64.35	64.92
mid α_1	64.90	65.44	65.06	62.46	63.93	64.74	64.75
sprs	47.48	47.70	50.36	48.18	47.76	45.46	48.01
$\{$mid $\alpha_1 \cup$ sprs$\}$	76.34	77.00	75.04	78.57	77.82	77.04	78.93
comp	76.22	77.00	70.23	78.65	77.50	76.65	78.19
alr	67.74	79.24	79.00	78.25	77.49	71.84	77.72
clr	67.69	79.24	79.53	78.09	77.82	68.23	77.08
ilr	67.69	79.24	79.53	78.09	77.82	68.23	77.08

Table 5.52: *Ling Sex* geometric mean classification in Fuzzy space.

BCCF1	BCCF2	BCCF3	BCCF4
65.73	70.27	73.13	77.58

6
Epilogue

6.1 Final remarks

The results yielded by the fuzzy approach when algorithms are given some freedom turned out to be very good. The fuzzy algorithm leads to a conceptually lazy model that is easy to understand. The behavior of the compositional approach has also led to interesting results. On the other hand, for some of the classical algorithms and the direct multivariate approach are not adequate for general settings, as they are unable to build features such as variance.

6.2 Summary of contributions

Two supervised classification problems were evaluated against a batch of classifiers in a naive multivariate way, and in a novel simplex compositional data approach which is compatible with the kind of data generated by means of fuzzy convex constrained trapezoids. The classification problem was also tested in a fuzzy theoretical framework against two conventional classifiers and two novel ones, which are based upon the initial definitions of ball based fuzzy ones. The last fuzzy classifier we proposed presents an improvement over the previous ones, and performs quite well. Besides, it's robust problem wise against more classical approaches.

6.3 Conclusions and future work

This work is an initial study concerning the problem of supervised classification of random fuzzy sets. We proposed to use two additional approaches to better capture the generality one can encounter in a classification problem in $\mathcal{F}_c(\mathbb{R})$—i.e. adjusting by location and/or dispersion. The fuzzy trapezoidal convex data treated as compositional data performs quite well. After this initial proposal it could interesting to see if this *ad-hoc* procedure could be pursued any further under the fuzzy statistics umbrella.

Bibliography

J. Aitchison. *The statistical analysis of compositional data (2003 reprint)*. Blackburn Press, 1986.

I.E. Allen and C.A. Seaman. Likert scales and data analyses. *Quality Progress*, 40(7):64–65, 2007.

M Ángeles Gil. Fuzzy statistics and imprecise probabilities. course slides from the University of Oviedo Master in Soft Computing and Intelligent Data Analysis, 2010.

A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn. Compositional data analysis in the geosciences. *Geological Society Special Publication*, 204, 2007.

C. Castaing, M. Valadier, and SpringerLink (Service en ligne). *Convex analysis and measurable multifunctions*, volume 580. Springer-Verlag Berlin, 1977.

A. Colubi. Statistical inference about the means of fuzzy random variables: Applications to the analysis of fuzzy-and real-valued data. *Fuzzy Sets and Systems*, 160(3):344–356, 2009.

A. Colubi and G. González-Rodríguez. Triangular fuzzification of random variables and power of distribution tests: Empirical discussion. *Computational statistics & data analysis*, 51(9):4742–4750, 2007.

A. Colubi, G. González-Rodríguez, M. Ángeles Gil, and W. Trutschnig. Nonparametric criteria for supervised classification of fuzzy data. *International Journal of Approximate Reasoning*, 2011.

David B. Dahl. *xtable: Export tables to LaTeX or HTML*, 2012. URL http://CRAN.R-project.org/package=xtable. R package version 1.7-0.

J.J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003.

G. González-Rodríguez, A. Colubi, and M.Á. Gil. Fuzzy data treated as functional data: a one-way anova test approach. *Computational Statistics & Data Analysis*, 2010.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

R. Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

M.L. Puri and D.A. Ralescu. The concept of normality for fuzzy random variables. *The Annals of Probability*, 13(4):1373–1379, 1985.

M.L. Puri and D.A. Ralescu. Fuzzy random variables. *Journal of Mathematical Analysis and Applications*, 114(2):409–422, 1986.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org/. ISBN 3-900051-07-0.

Charlie Sharpsteen and Cameron Bracken. *tikzDevice: A Device for R Graphics Output in PGF/TikZ Format*, 2012. URL http://CRAN.R-project.org/package=tikzDevice. R package version 0.6.2.

W. Trutschnig, G. González-Rodríguez, A. Colubi, and M.Á. Gil. A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. *Information Sciences*, 179(23):3964–3972, 2009.

Wolfgang Trutschnig and Asun Lubiano. *SAFD: Statistical Analysis of Fuzzy Data*, 2011. URL http://CRAN.R-project.org/package=SAFD. R package version 0.3.

K. Gerald van den Boogaart, Raimon Tolosana, and Matevz Bren. *compositions: Compositional Data Analysis*, 2011. URL http://CRAN.R-project.org/package=compositions. R package version 1.10-2.

L.A. Zadeh. The concept of a linguistic variable and its application to approximate reasoning–I. *Information sciences*, 8(3):199–249, 1975.