UNIVERSITÀ DEGLI STUDI DI PALERMO
DIPARTIMENTO SCIENZE ECONOMICHE, AZIENDALI E STATISTICHE
SECS-S/01 - STATISTICAL SCIENCE

DOCTORAL THESIS

in

Scienze Economiche e Statistiche
Ciclo XXXII

Ensemble methods for ranking data with and without position weights

---

Ph.D. candidate:
Simona BUSCEMI

Supervisor:
Prof. Antonella PLAIA

Co-ordinator:
Prof. Andrea CONSIGLIO

Co-supervisor:
Prof. Gianfranco LOVISON

"First things first, but not necessarily in that order."

[Doctor Who]

"Chaos was the law of nature. Order was the dream of man."

[Henry B. Adams]

Università degli Studi di Palermo

Abstract

Ensemble methods for ranking data with and without position weights

by Simona BUSCEMI

The main goal of this Thesis is to build suitable Ensemble Methods for ranking data with weights assigned to the items'positions, in the cases of rankings with and without ties.

The Thesis begins with the definition of a new rank correlation coefficient, able to take into account the importance of items'position. Inspired by the rank correlation coefficient, τ_x, proposed by Emond and Mason (2002) for unweighted rankings and the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010), this work proposes τ_x^w, a new rank correlation coefficient corresponding to the weighted Kemeny distance. The new coefficient is analized analitically and empirically and represents the main core of the consensus ranking process. Simulations and applications to real cases are presented. In a second step, in order to detect which predictors better explain a phenomenon, the Thesis proposes decision trees for ranking data with and without weights, discussing and comparing the results. A simulation study is built up, showing the impact of different structures of weights on the ability of decision trees to describe data. In the third part, ensemble methods for ranking data, more specifically Bagging and Boosting, are introduced. Last but not least, a review on a different topic is inserted in this Thesis. The review compares a significant number of linear mixed model selection procedures available in the literature. The review represents the answer to a pressing issue in the framework of LMMs: how to identify the best approach to adopt in a specific case. The work outlines mainly all approaches found in literature. This review represents my first academic training in making research.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisor Prof. Antonella Plaia for her continuous support of my Ph.D study and research, for her patience, expertise, motivation and enthusiasm. Her guidance helped me in all the time of my research and writing this thesis. I could not have imagined having a better advisor and mentor for my Ph.D studies.

Besides my advisor, I would like to thank the KE group of Darmstadt University: Prof. J. Fuernkranz and the Post Doc Eneldo Loza Mencia, for offering me the abroad internship in their group with encouragement, insightful comments and hard questions. I also would like to thank them for having allowed me to work with the HLR cluster system, which had a predominant role for running the R codes.

My gratitude also goes to Prof. Antonio D'Ambrosio for his huge experience and his really useful suggestions about the topic of my thesis, during many conferences and dinners, we had together.

I would like to extend my sincere gratitude to Prof. Andrea Consiglio, the PhD activities'coordinator, and all the members of the Department SEAS from Palermo University, but in particular to Mariangela Sciandra for her academic contribution and for her humanity, Gianluca Sottile for his academic contribution that helped me in dark moments, Prof. Gianfranco Lovison and his wife Marcella for the many stimulating talks we had, Giovanni Boscaino for being so funny and motivating person meanwhile, Giada Adelfio who is such a sensitive person, Vito Muggeo and Prof. Marcello Chiodi that are definitely one of a kind, Antonio Abruzzo and Luigi Augugliaro for sharing funny moments in the department during launch times.

I'm taking with me unforgettable memories together with my lovely colleagues, therefore a sincere thank you goes to many of them: Simona for our friendship and unconditional and shared love for cats, Martina for her "cuore di panna", Cinzia, Chiara, Salvo C., Salvo P., Rodolfo and Jelena for laughing so much all together even if we lived so stressful time together.

I thank all my friends (Francesca, Ornella, Adriana, Johana, Roberta, Andrea, Chiara, Jenny, Salvo, Lidia, Federico, Valentina, Merilin and many others) and everyone who helped me to appreciate life while i was working on this project and to all those friends I met in Darmstadt (Oliver, Farina, Tobias, Jonas and Aicha) for keeping me company on long walks and funny dinners.

I should like to express my heartfelt thanks to Giusy U. and Maria Laura S.: thank you for the enlightenment about many aspects of my life.

A heartful thank you goes to Sandra, and her beautiful daughter, for having shared with me many touching experiences in these years.

Last but not least, I would like to thank my family: my mom for being my wonderful mom, my special aunt Vita and his husband (Zio Peppino), my cousins Graziella, Emma and Sandro, Antonella, Mauro and Marilena, my aunt Alfonsina and my uncle Gaetano... for supporting me spiritually throughout my life.

This thesis is dedicated to you, dad, supervising me from unknown places and always in my heart. I love you and I miss you so much.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Ranking data: weights to items'positions

Ranking and classification are basic cognitive ways that people generally use for grading everything they experience. Grouping and ordering a set of elements (movies, politicians, sport teams, websites, and so on) represent a natural and common attitude for human beings in handling their life. A particular kind of ranking data is given by preferences, where individuals (called "judges") provide their preferences over a set of objects (called "items"). The traditional metrics comparing rankings don't take into account the importance of swapping elements similar among them (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights) which are a crucial point in some fields of research. First of all, this Thesis focuses on the importance given to items'positions.
Within preference data framework, distance-based decision trees represent a non-parametric tool for identifying the profiles of subjects giving a similar ranking. This Thesis aims to detect, in the framework of complete ranking data, with or without ties, the impact of different structures of weighted distances for, firstly, building decision trees and, after that, Ensemble methods.

1.2 Distance-based models for weighted ranking data

The last years have seen a remarkable flowering of works about the use of decision trees.
Decision trees are non parametric recursive statistical tools used for classification and prediction issues. The so-called decision trees are named in such a way because of their tree-shapped structure, obtained by some prediction rules. The most known decision tree methodology consists of

classification and regression tree (CART, Breiman *et al.* (1984)), and it is based on two steps: growing and pruning. CART methodology has been introduced for predicting quantitative or categorical variables and, only, in few cases for analyzing ordinal responses or rankings.
As a matter of fact, decision trees are useful and intuitive, but they are very unstable: small perturbations bring big changes. For this reason it's necessary to use more stable procedures, as ensemble methods, in order to find which predictors are able to explain the preference structure in a more efficient way. In this work ensemble methods as Bagging and Boosting are proposed, from both a theoretical and computational point of view, for deriving multiple classification trees when ranking data (with and without position weights) are observed. The advantages of these procedures are shown through an example and a simulation. The last topic (selection of effects in Linear Mixed Models) could seem far away from the consensus ranking problem, but, actually, we could consider the output of a model selection process as a ranking; therefore, using different measures (AIC, BIC,...) which provide different rankings of the models, a consensus ranking process could be applied in order to identify the "optimum"ranking of the models. In a few words, each measure could provide a ranking of linear mixed models and a consensus ranking process could be useful for detecting the best consensus of the ranked models.

1.3 Outline of thesis

The contribution of this thesis involves the development of a new rank correlation coefficient for ranking (complete and weak) data, that can be used as the main ingredient of consensus measure processes in order to cope with the rank aggregation (in the terminal nodes of distance-based decision trees and, suitably, in the aggregation procedure of predicted rankings in the field of Ensemble Methods). As final chapter a review about the selection of effects in the framework of linear mixed models. A summary outline of how the thesis is organized can be shown:

- **Foundations on rankings**. Chapter 2 forms the basis for all subsequent chapters. It describes the main classical distances used for rankings and the correlation coefficients related to each distance.

- **A new position weighted rank correlation coefficient without ties**. In this thesis, a new position weighted correlation coefficient for

consensus ranking process is provided, able to deal with linear rankings. Chapter 3 has been published as (Plaia *et al.*, 2019b) and it proposes a new rank correlation coefficient when ties are not allowed.

- **A new position weighted rank correlation coefficient with ties**. The new position weighted correlation coefficient for consensus ranking process is adapted for weak rankings. Chapter 4 has been published as a conference paper (Plaia *et al.*, 2018b) and it is actually submitted (Plaia *et al.*, 2018a).

- **Decision trees for weighted ranking data.** Chapter 5 proposes decision trees for positional weighted ranking data, using the new rank correlation coefficient introduced in the previous chapter. This chapter is an extension of the short paper submitted at the conference SIS2018 (Plaia *et al.*, 2018c).

- **Boosting and Bagging for ranking data, with and without position weights.** In Chapter 6, multiple decision trees are combined for rankings with position weights, leading to the development of suitable Ensemble Methods, such as Bagging and Boosting. Chapter 6 is an extension of a paper submitted to the conference ASMDA2019 (Plaia *et al.*, 2019a).

- **Model Selection in Linear Mixed-Effect Models: a Review.** Chapter 7 ends the Thesis focusing on a different topic: model selection in Linear Mixed-Effect Models (LMMs). Many approaches are studied and compared, focusing on the part of the model subject to selection (fixed and/or random), the dimensionality of models and the structure of variance and covariance matrices, and also, wherever possible, the existence of an implemented application of the methodologies set out. This chapter has been published as Buscemi and Plaia (2019).

Chapter 2

Foundations on rankings

2.1 Introduction

Distances between rankings and the rank aggregation problem have received growing consideration in the past few years. Ranking is one of the most simplified cognitive processes that help people to handle many aspects of their life. When some subjects are asked to indicate their preferences over a set of alternatives (items), ranking data are called preference data. An important issue involving rankings concerns the aggregation of the preferences in order to identify a compromise or a "consensus". Different approaches have been proposed in the literature to cope with this problem, but probably the most popular is the one related to distances/correlations. In order to get homogeneous groups of subjects having similar preferences, it is natural to measure the spread between rankings through dissimilarity or distance measures among them. In this sense, a consensus is defined to be the ranking that is the closest (i.e. it shows the minimum distance) to the whole set of preferences. Another possible way for measuring (dis)-agreement between rankings is in terms of a correlation coefficient: rankings in full agreement are assigned a correlation of +1, those in full disagreement are assigned a correlation of -1, and all others lie in between. A distance d between two rankings, instead, is a non-negative value, ranging in 0 - Dmax , where 0 is the distance between a ranking and itself. A distance measure d can be transformed into a correlation coefficient c (and vice-versa) using the linear transformation $c = 1 - \frac{2d}{D_{max}}$. Distances between rankings have received a growing consideration in the past few years. Usual examples of metrics in this framework are Kendall's and Spearman's. In 1962 Kemeny and Snell (1962) introduced a metric defined on linear and weak orders, known as Kemeny distance (or metric), which satisfies the constraints of a distance measure suitable for rankings. Cook (2006) highlights the intractability of the Kemeny metric, an issue already stressed by Emond

and Mason (2002): that's why the latter introduced a new correlation coefficient strictly related to Kemeny distance, proposing to use this coefficient in place of Kemeny metric, as bases for deriving a consensus among a set of rankings.

Even if nowadays the problem of the difficulty to cope with the Kemeny distances (due to the presence of absolute values) is behind us, thanks to its improvements, working with correlations rather than with distances is preferable due to its range, always between -1 and 1 independently on the particular distance used. On the contrary, D_{max} depends on the chosen distance.

The traditional metrics between rankings don't take into account the importance of swapping similar elements (element weights) or elements belonging to the top (or to the bottom) of an ordering (position weights). Kumar and Vassilvitskii (2010) have provided an extended measure for Spearman's Footrule (Spearman, 1987) and Kendall's Tau (Kendall, 1938), embedding weights relevant to the elements or to their position in the ordering. As Henzgen and Hüllermeier (2015) say, weighted versions of rank correlation measures have been studied in many fields other than statistics. For example, "in information retrieval, important documents are supposed to appear in the top, and a swap of important documents should incur a higher penalty than a swap of unimportant ones". In the context of the web, comparing the query results from the different search engines, the distance should emphasize the difference of the top elements more than the bottom ones, since people may be interested in the first few items(Chen *et al.*, 2014). A short review of the solutions proposed in the literature to cope with this issue can be found in Yilmaz *et al.* (2008).

The first purpose of this thesis is to propose a new position weighted correlation coefficient and to investigate the effect of different weighting vectors on the consensus ranking process. Particular attention is given to the weighted Kemeny distance (proposed by García-Lapresta and Pérez-Román (2010)) and a properly modified τ_x of Emond and Mason (2002) is defined, in order to measure the correlation between position weighted rankings. The choice of a rank correlation coefficient based on the Kemeny distance is due to the ability of τ_x to measure the correlation between ranks with ties. Even if in the first part of this work the focus is on rankings without ties, this can be considered the first step for the definition of a generalized coefficient able to capture the correlation giving the right importance to both the positions of items and to ties.

2.2 Distances between rankings and rank correlation coefficients

2.2.1 Classical distances for ranking data

Ranking data arise when a group of n judges (experts, voters, raters etc.) is asked to rank a fixed set of m *items* (different alternatives of objects like movies, activities and so on) according to their preferences.

While ranking m items, labelled $1, \dots m$, a ranking a is a mapping function from the set of items $\{1, \dots, m\}$ to the set of ranks $\{1, \dots, m\}$, endowed with the natural ordering of integers, where a_i is the rank given by the judge to item i [1]. When all m items are ranked in m distinct ranks, we observe a complete ranking or *linear ordering* (Cook *et al.*, 1986). A ranking a is, therefore, one of the $m!$ possible permutations of m elements, containing the preferences of a judge for the m items. Yet, it is also possible that a judge fails to distinguish between two or more objects and assigns them equally, thus resulting in a tied ranking or *weak ordering*. In real situations, many times it happens that not all items are ranked and, so, besides complete and tied rankings, *partial* and *incomplete rankings* exist: the first occurs when only a specific subset of $q < m$ objects are ranked by judges, while incomplete rankings occur when judges are free to rank different subsets of m objects (Cook *et al.*, 1986). Obviously, different types of ordering will generate different sample spaces of ranking data. With m objects there are $m!$ possible complete rankings; this number gets even larger when ties are allowed (for the cardinality of the universe when ties are allowed refer to Good (1980) and Marcus *et al.* (2013).

In order to classify judges into C homogeneous clusters according to their expressed preferences, a dissimilarity or distance measure d has to be defined for rankings and such a measure has to meet the usual properties of a distance function:

- Reflexivity: $d(a, a) = 0$,
- Positivity: $d(a, b) > 0$ if $a \neq b$,
- Symmetry: $d(a, b) = d(b, a)$,
- Triangle inequality: $d(a, b) \leq d(a, c) + d(c, b)$ (in case of a distance).

[1] Preference rankings can be represented through either rank vectors (as in this chapter) or order vectors (D'ambrosio *et al.*, 2015a)

Moreover, a desirable property of any distance is its invariance toward a renumbering of the elements: the so-called label invariance or equivariance (Cheng *et al.*, 2009).

Many distances between rankings can be found in the literature (Marcus *et al.*, 2013), given in terms of ranking themselves, or as functions of the $m \times m$ score matrix $\{a_{ij}\}$ (defined for the generic ranking a), whose elements are defined as:

$$a_{ij} = \begin{cases} 1 & \text{if object } i \text{ is preferred to object } j \\ -1 & \text{if object } j \text{ is preferred to object } i \\ 0 & \text{if objects } i \text{ and } j \text{ are tied, or if } i = j \end{cases} \tag{2.1}$$

Let's see a small example that shows how to compute the score matrix. Given the ranking vectors $a = (3, 1, 1, 4)$ and $b = (1, 3, 2, 2)$ The score matrices of a and b, computed according to (2.1), are respectively:

$$a = \begin{bmatrix} 0 & -1 & -1 & +1 \\ & 0 & 0 & +1 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix} \quad b = \begin{bmatrix} 0 & +1 & +1 & +1 \\ & 0 & -1 & -1 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}$$

Another possible way of measuring (dis)-agreement between rankings is in terms of a correlation coefficient: rankings in full agreement are assigned a correlation of $+1$, those in full disagreement are assigned a correlation of -1, and all others lie in between. A distance between two rankings, instead, is a non-negative value, ranging in $0 - D_{max}$, where 0 is the distance between a ranking and itself. A distance measure d can be transformed into a correlation coefficient c (and vice-versa) using the linear transformation $c = 1 - \frac{2\,d}{D_{max}}$ (Emond and Mason, 2002).

One of the best-known metrics that evaluate the distance between two permutations is *Spearman's Footrule* (Spearman, 1987) distance, which measures the ℓ_1 distance between two generic orderings a and b as follows:

$$F(a, b) = \sum_{i=1}^{m} |a_i - b_i|. \tag{2.2}$$

As an alternative, *Kendall's Tau* (Kendall, 1938) distance is defined as the number of discordant pairs between rankings a and b:

$$T(a, b) = \sum_{1 \le i} \sum_{i < j} I\{[a_i - a_j][b_i - b_j] < 0\}, \tag{2.3}$$

where I is the indicator function. Its expression in terms of correlation, τ_b, can be found in the literature as an extension of *Kendall's Tau* rank correlation coefficient to the case of weak orderings, by using the score matrices a_{ij} and b_{ij} of the two rankings a and b defined according to Equation 2.1:

$$\tau_b(a,b) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}b_{ij}}{\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}^2 \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij}^2}}. \tag{2.4}$$

In the case of linear orderings comparison, the denominator of τ_b reduces to $m(m-1)$; when weak orderings are compared, the denominator assumes a smaller value, reduced according to the total number of ties observed for each ranking (Emond and Mason, 2002).

Kemeny and Snell (1962) outlined a set of four axioms that should apply to a distance measure between weak orderings, and proposed a new distance, also defined in terms of score matrices, that satisfies these axioms:

$$K(a,b) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} |a_{ij} - b_{ij}| \tag{2.5}$$

The Kemeny distance, $K(a,b)$, between two rankings a and b is a city-block distance where a_{ij} and b_{ij} are the generic elements of the $m \times m$ score matrices associated to a and b.

Emond and Mason (2000) showed that Spearman's Footrule distance in Equation (2.2) suffers from what is known as sensitivity to irrelevant *items*, that can lead to an inconsistent result. The same authors (Emond and Mason, 2002) proved later that Kendall's distance failed the triangular inequality when dealing with weak orderings; therefore, they proposed a new correlation coefficient τ_x, that differs from Kendall's τ_b by using a different score matrix $\left\{a'_{ij}\right\}$ to represent ties:

$$a'_{ij} = \begin{cases} 1 & \text{if object } i \text{ is preferred to or tied with object } j \\ -1 & \text{if object } j \text{ is preferred to object } i \\ 0 & \text{if objects } i \text{ and } j \text{ are the same} \end{cases} \tag{2.6}$$

Let's consider the small example that shows how to compute the score matrix, according to (2.6). Given the ranking vectors $a = (3,1,1,4)$ and

$b = (1, 3, 2, 2)$ The score matrices of a and b, defined by Emond and Mason (2002), are respectively:

$$a' = \begin{bmatrix} 0 & -1 & -1 & +1 \\ & 0 & +1 & +1 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix} \quad b' = \begin{bmatrix} 0 & +1 & +1 & +1 \\ & 0 & -1 & -1 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix}$$

The new rank correlation coefficient, τ_x, is defined as:

$$\tau_x(a, b) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{m} a'_{ij} b'_{ij}}{m(m-1)}. \tag{2.7}$$

τ_x reduces to τ_b for linear orders (i.e. in the absence of ties) and the authors demonstrate that it is the correlation coefficient corresponding to the Kemeny's distance in Equation (2.5).

2.2.2 Weighted distances

The distances presented in Section (2.2.1) fail to take into account two important aspects: *element* and *positional* information (Kumar and Vassilvitskii, 2010). In some practical applications, one or some of the k elements can be more important than others, or, similarly, the top of the ordering can deserve more attention than the bottom. In these situations, changing the rank of very important elements or changing the top of the ranking requires different "weighting". Lee and Yu (2010) proposed a distance-based tree model where weights are functions related to modal ranking.

In this thesis, on the other hand, we focus on position weights, while considering the weighted version of the three distances (F, T, and K) presented in the previous section. Nevertheless, even by introducing position weights, the maximum distance between two rankings is reached when one ranking is the exact reverse of the other.

Let $w = (w_1, w_2, \ldots, w_{m-1})$ be a weighting vector, such that $\sum_{i=1}^{m-1} w_p = 1$ and $w_1 \geq w_2 \geq \ldots \geq w_{m-1} > 0$; weight w_i is used to measure the contribution of moving an element from position i to position $i+1$ or from position $i+1$ to position i to the overall distance. Thus, for any two positions i and j, $C_F(i, j)$, defined as:

$$C_F(i, j) = \begin{cases} \sum_{k=i}^{j-1} w_k & \text{if } i < j \\ 0 & \text{if } i = j \\ \sum_{k=j}^{i-1} w_k & \text{if } i > j \end{cases}$$

will be the total contribution of moving an element from position i to position j. For permutations a and b, the contribution of moving a generic element i is $C_F(a(i), b(i))$ and the resulting *Weighted Spearman's Footrule* distance (Chen *et al.*, 2012) will be defined as:

$$F^w(a,b) = \sum_{i=1}^{k} C_F(a(i), b(i)) \tag{2.8}$$

An extension to the weighted case of *Kendall's Tau* distance in Equation (2.3) is due to Farnoud *et al.* (2012). It is based on the intuitive concept that given two rankings a and b, it is always possible to obtain a from b by a sequence of adjacent inversions. So let $s = (\eta_1, \eta_2, \ldots, \eta_m)$ be a *transforming sequence* of consecutively executed permutations where each one inverts two adjacent ranks, and that transforms a to b and $C_T(s) = \sum_{r=1}^{m} w_{\eta_r}$ the *length* of the transforming sequence s, then *Weighted Kendall's Tau* distance between a and b is defined as:

$$T^w(a,b) = \min_{s \in \Phi} C_T(s) \tag{2.9}$$

over the set Φ of all possible transforming sequences from a to b.

In this part of the thesis, we consider the weighted version of the Kemeny metric proposed by García-Lapresta and Pérez-Román (2010) (see Equation (2.10)) and we limit the analyses to the case of linear orders. For measuring the weighted distances, the non-increasing weighting vector $w = (w_1, w_2, \ldots, w_{m-1})$ is used, where w_i is the weight given to position i in the ranking, with $\sum_{i=1}^{m-1} w_i = 1$. Note that w_i is the weight that *we want* to give to position i, and is not derived from the data, but chosen by the analyst. Given two generic rankings of m elements, a and b, the Weighted Kemeny distance was provided by García-Lapresta and Pérez-Román (2010) as follows:

$$d_K^w(a,b) = \frac{1}{2}\left[\sum_{\substack{i,j=1 \\ i<j}}^{m} w_i |a_{ij}^{(\sigma_1)} - b_{ij}^{(\sigma_1)}| + \sum_{\substack{i,j=1 \\ i<j}}^{m} w_i |b_{ij}^{(\sigma_2)} - a_{ij}^{(\sigma_2)}|\right], \tag{2.10}$$

where (σ_1) states to follow the a ranking and (σ_2), similarly, orders according to b. More specifically, $b_{ij}^{(\sigma_1)}$ is the score matrix of the ranking b reordered according to a, $a_{ij}^{(\sigma_2)}$ is the score matrix of the ranking a reordered according to b and $a_{ij}^{(\sigma_1)} = b_{ij}^{(\sigma_2)}$ is the score matrix of the linear order $1, 2, \ldots, m$ (see Plaia and Sciandra (2019) for more details).

A small example shows how to compute the weighted Kemeny distance. Given the ranking vectors $a = (3, 1, 1, 4)$ and $b = (1, 3, 2, 2)$, we can easily define their orderings: $\sigma_1 = (2, 3, 1, 4)$ and $\sigma_2 = (1, 3, 4, 2)$. Then we have: $a^{\sigma_1} = (1, 1, 3, 4)$, $a^{\sigma_2} = (3, 1, 4, 1)$, $b^{\sigma_1} = (3, 2, 1, 2)$ and $b^{\sigma_2} = (1, 2, 2, 3)$. The score matrices of a^{σ_1}, b^{σ_1}, a^{σ_2} and b^{σ_2} are respectively:

$$a^{\sigma_1} = \begin{bmatrix} 0 & 0 & +1 & +1 \\ & 0 & +1 & +1 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix} \quad b^{\sigma_1} = \begin{bmatrix} 0 & -1 & -1 & -1 \\ & 0 & -1 & 0 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix}$$

$$b^{\sigma_2} = \begin{bmatrix} 0 & +1 & +1 & +1 \\ & 0 & 0 & +1 \\ & & 0 & +1 \\ & & & 0 \end{bmatrix} \quad a^{\sigma_2} = \begin{bmatrix} 0 & -1 & +1 & -1 \\ & 0 & +1 & 0 \\ & & 0 & -1 \\ & & & 0 \end{bmatrix}$$

Through simple algebric operations it can be easily found that:

$$|a^{\sigma_1} - b^{\sigma_1}| = \begin{bmatrix} 0 & 1 & 2 & 2 \\ & 0 & 2 & 1 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix} \quad |b^{\sigma_2} - a^{\sigma_2}| = \begin{bmatrix} 0 & 2 & 0 & 2 \\ & 0 & 1 & 1 \\ & & 0 & 2 \\ & & & 0 \end{bmatrix}$$

Now we can compute the weighted Kemeny distance. Let's assume equal weights, $w = (1/3, 1/3, 1/3)$, the distance between a and b is:

$$d_K^w(a, b) = \frac{1}{2}\left[(1 + 2 + 2 + 2 + 0 + 2)\frac{1}{3} + (2 + 1 + 1 + 1)\frac{1}{3} + (0 + 2)\frac{1}{3}\right] =$$

$$= \frac{1}{2}\left(\frac{16}{3}\right) = \frac{8}{3}.$$

If we change the vector of weights, with a decreasing structure $w = (2/3, 1/3, 0)$, the value of the distance is:

$$d_K^w(a, b) = \frac{1}{2}\left[(1 + 2 + 2 + 2 + 0 + 2)\frac{2}{3} + (2 + 1 + 1 + 1)\frac{1}{3} + (0 + 2)0\right] =$$

$$= \frac{1}{2}\left(\frac{23}{3}\right) = \frac{23}{6}.$$

For the sake of completeness, other kinds of weight could be introduced when considering preference data: weights assigned to individuals. Hence, a weight is not assigned to an item or a position, but to

the whole ranking, i.e. to the judge. This necessity can arise when dealing with measures of consensus ranking, and a weight w_j, assigned to individual j, represents the strength of his opinion among the group of individuals (Emond and Mason, 2002).

Chapter 3

A position weighted rank coefficient for rankings without ties

As already stated in Section 2.1, Cook (2006) highlights the intractability of the Kemeny metric, issue already stressed by Emond and Mason (2002). Given a set of n independent rankings involving m items, the median ranking is the one for which the Kemeney-Snell distance is the minimum. According to this approach, researching the median ranking implies to search the space of all possible rankings with m objects. The research is totally governed by the number of items and not by the number of judges. In the case of full rankings the set of all different rankings Z^m is equal to $m!$. The searching space increases consistently when ties are allowed: the universe of all possible rankings with ties (S^m) is approximately given by the following quantity:

$$S^m = \sum_{r=0}^{m} r! \left\{ {m \atop r} \right\},$$

where $\left\{ {m \atop r} \right\}$ states the Stirling number of the second kind, corresponding to the number of all possible ways to partition a set of m objects into r non-empty subsets. "Even if there are close formulas for the detection of the consensus rankings, these are not feseable because of the complexity of the problem (e.g. when ties are allowed , in the case of $m = 12$ we have that $S^n = 28.091.567.595$)"(D'Ambrosio *et al.*, 2017). That's a NP hard problem, because of the high values reached by the Stirling number of the second type when the number of items increases and ties are allowed. That's the reason why the definition of a rank correlation coefficient (with positional weights) has been necessary. Moreover, in statistics a measure of correlation is, usually, an index assuming values in $[-1, 1]$ which gives

an information deeper than only a distance, i.e. how much is the intensity of the relation between two variables. Hence, working with correlations rather than with distances is preferable due to its range, always between -1 and 1 independently on the particular distance used. On the contrary, the maximum value of a distance,D_{max}, depends on the chosen distance.

3.1 Linear orderings

In this work we propose a new rank correlation coefficient, suitable for position weighted rankings which handles linear orders.

Considering that, even if Emond and Mason (2002) introduced their τ_x as in Equation (2.7), it can also be written as:

$$\tau_x^w(a,b) = \frac{\sum_{i<j}^m a_{ij}^{(\sigma_1)} b_{ij}^{(\sigma_1)} + \sum_{i<j}^m a_{ij}^{(\sigma_2)} b_{ij}^{(\sigma_2)}}{m(m-1)},$$

combining the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010) and the extension of τ_x provided by Emond and Mason (2002), we define:

$$\tau_x^w(a,b) = \frac{\sum_{i<j}^m a_{ij}^{(\sigma_1)} b_{ij}^{(\sigma_1)} w_i + \sum_{i<j}^m a_{ij}^{(\sigma_2)} b_{ij}^{(\sigma_2)} w_i}{\max[d_K^w(a,b)]}, \tag{3.1}$$

where the denominator represents the maximum value for the Kemeny weighted distances (García-Lapresta and Pérez-Román, 2010), equal to:

$$\max[d_K^w(a,b)] = 2\sum_{i=1}^{m-1}(m-i)w_i. \tag{3.2}$$

It can be proven that the constraint $\sum_{i=1}^{m-1} w_i = 1$ is no more necessary. For linear orderings (i.e. without ties), a^{σ_1} and b^{σ_2} represent the natural ascending orderings. Let $a^{\sigma_1} = b^{\sigma_2} = 1, 2, \ldots, m$, the new rank correlation coefficient reduces to:

$$\tau_x^w(a,b) = \frac{\sum_{i<j} A_{ij}(b_{ij}^{(\sigma 1)} + a_{ij}^{(\sigma 2)})w_i}{\max[d_K^w(a,b)]},$$

where

$$A_{ij} = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 & 1 \\ -1 & 0 & 1 & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \dots & -1 & 0 \end{bmatrix}$$

3.2 Correspondance between distance and correlation

The correspondence between the weighted rank correlation coefficient and the weighted Kemeny distance holds, whatever is the weighting vector assigned to the items' positions:

$$\tau_x^w = 1 - \frac{2d_k^w}{\max(d_k^w)}$$

or equivalently, it's enough to demonstrate the equality of the following equation:

$$\sum_{i<j}^{m} A_{ij} \left(b_{ij}^{(\sigma^1)} + a_{ij}^{(\sigma^2)} \right) w_i =$$

$$= 2 \sum_{i=1}^{m-1} (m-i) w_i - \left[\sum_{i<j}^{m} w_i |A_{ij} - b_{ij}^{(\sigma_1)}| + \sum_{i<j}^{m} w_i |A_{ij} - a_{ij}^{(\sigma_2)}| \right]. \quad (3.3)$$

Proof. In order to evaluate the contribution to the sum in Equation (3.2) we can distinguish the two cases:

Case 1. Both A and B prefer object i to j. The Kemeny-Snell matrix values are: $A_{ij} = b_{ij}^{(\sigma_1)} = a_{ij}^{(\sigma_2)} = 1$.
The τ_x^w score matrix values are: $a_{ij}^{'(\sigma_2)} = b_{ij}^{'(\sigma_1)} = 1$ Hence, the equality in Equation (3.2) holds:
$1 + 1 = 2 - [|1 - (1)| + |1 - (1)|]$.

Case 2. A prefers object i to j and B prefers j to object i. The Kemeny-Snell matrix values are: $A_{ij} = 1$ and $b_{ij}^{(\sigma_1)} = a_{ij}^{(\sigma_2)} = -1$.
The τ_x^w score matrix values are: $a_{ij}^{'(\sigma_2)} = b_{ij}^{'(\sigma_1)} = -1$ Hence, the equality in Equation (3.2) holds:
$-1 - 1 = 2 - [|1 - (-1)| + |1 - (-1)|]$.

Proof. The equivalence can be also demonstrated considering the concordant and discordant couples (analogous approach is used in Vigna (2015)),

$$\underbrace{\sum_{i<j}^{m} A_{ij}\left(b_{ij}^{(\sigma^1)}+a_{ij}^{(\sigma^2)}\right)w_i}_{(1)^*} = \underbrace{2\sum_{i=1}^{m-1}(m-i)w_i}_{(2)^*} - \underbrace{\left[\sum_{i<j}^{m} w_i|A_{ij}-b_{ij}^{(\sigma_1)}|+\sum_{i<j}^{m} w_i|A_{ij}-a_{ij}^{(\sigma_2)}|\right]}_{(3)^*}.$$

Fixing i and w_i, each contribution to the sums in $(1)^*$, $(2)^*$ and $(3)^*$ is given by:
$(1)^*$ $\sum_{j=i+1}^{m} A_{ij}(b_{ij}^{(\sigma^1)}+a_{ij}^{(\sigma^2)}) = c_i - d_i$.
where c_i and d_i are the number of all concordant and discordant couples of items for the two rankings a and b, for the i-th item, respectively (see the following example).
Since $\sum_{i=1}^{m-1}(m-i)w_i = \sum_{i<j} w_i$, fixing i and w_i, each piece of the sum in $(2)^*$ is:
$(2)^*$ $2\sum_{j=i+1}^{m} 1 = 2(m-i) = c_i + d_i$, i.e. the number of all couples of items for two rankings, row by row.
As regards the last expression $(3)^*$, each single element of the sum is:
$(3)^*$ $\sum_{j=i+1}^{m}(|A_{ij}-b_{ij}^{(\sigma_1)}|+|A_{ij}-a_{ij}^{(\sigma_2)}|) = 2[2(m-i)-c_i] = 2d_i$
Substituting row by row all the results in $(1)^*$, $(2)^*$ and $(3)^*$, hence fixing i and w_i, it can be easily seen that the equation in (3.2) holds:

$$c_i - d_i = c_i + d_i - 2d_i \Rightarrow c_i - d_i = c_i - d_i$$

Example: Let's consider $a = 1,2,3,4,5,6$ and $b = 2,4,1,6,5,3$. Hence, $a^{\sigma_1} = b^{\sigma_2} = 1,2,3,4,5,6$ and $a^{\sigma_2} = 3,1,6,2,5,4$ and $b^{\sigma_1} = 2,4,1,6,5,3$. The corresponding score matrices are:

$$a^{\sigma_1} = b^{\sigma_2} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ & 0 & 1 & 1 & 1 & 1 \\ & & 0 & 1 & 1 & 1 \\ & & & 0 & 1 & 1 \\ & & & & 0 & 1 \\ & & & & & 0 \end{bmatrix} \quad b^{\sigma_1} = \begin{bmatrix} 0 & 1 & -1 & 1 & 1 & 1 \\ & 0 & -1 & 1 & 1 & -1 \\ & & 0 & 1 & 1 & 1 \\ & & & 0 & -1 & -1 \\ & & & & 0 & -1 \\ & & & & & 0 \end{bmatrix}$$

$$a^{\sigma_2} = \begin{bmatrix} 0 & -1 & 1 & -1 & 1 & 1 \\ & 0 & 1 & 1 & 1 & 1 \\ & & 0 & -1 & -1 & -1 \\ & & & 0 & 1 & 1 \\ & & & & 0 & -1 \\ & & & & & 0 \end{bmatrix}$$

From the score matrices we are able to compute the concordant and discordant couples of items, row by row.

$$b^{\sigma_1} \Rightarrow \begin{array}{c|c} c_i & d_i \\ \hline 4 & 1 \\ 2 & 2 \\ 3 & 0 \\ 0 & 2 \\ 0 & 1 \end{array} \qquad a^{\sigma_2} \Rightarrow \begin{array}{c|c} c_i & d_i \\ \hline 3 & 2 \\ 4 & 0 \\ 0 & 3 \\ 2 & 0 \\ 0 & 1 \end{array}$$

The expression in (1), row by row and ignoring the presence of the weighting vector, is easily computed as follows:

$$b^{\sigma_1} + a^{\sigma_2} = \begin{bmatrix} 0 & 0 & 0 & 0 & 2 & 2 \\ & 0 & 0 & 2 & 2 & 0 \\ & & 0 & 0 & 0 & 0 \\ & & & 0 & 0 & 0 \\ & & & & 0 & -2 \\ & & & & & 0 \end{bmatrix}$$

$$\Rightarrow \quad \sum_{j=i+1} (b^{\sigma_1} + a^{\sigma_2}) = \begin{bmatrix} 4 \\ 4 \\ 0 \\ 0 \\ -2 \end{bmatrix} = \begin{bmatrix} (4+3)-(1+2) \\ (2+4)-(2+0) \\ (3+0)-(0+3) \\ (0+2)-(2+0) \\ (0+0)-(1+1) \end{bmatrix} = c_i - d_i$$

$$\underbrace{|A - b^{\sigma_1}|}_{2(m-i)-c_i(b^{\sigma_1})} + \overbrace{|A - a^{\sigma_2}|}^{2(m-i)-c_i(a^{\sigma_2})} =$$

$$= \begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ & 0 & 2 & 0 & 0 & 2 \\ & & 0 & 0 & 0 & 0 \\ & & & 0 & 2 & 2 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix} + \begin{bmatrix} 0 & 2 & 0 & 2 & 0 & 0 \\ & 0 & 0 & 0 & 0 & 0 \\ & & 0 & 2 & 2 & 2 \\ & & & 0 & 0 & 0 \\ & & & & 0 & 2 \\ & & & & & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 2(1+2) \\ 2(2+0) \\ 2(0+3) \\ 2(2+0) \\ 2(1+1) \\ 0 \end{bmatrix} = 2d_i$$

3.3 Minimum and Maximum values

The weighted rank correlation coefficient holds the main property of a generic correlation coefficient, i.e. it takes values between -1 and 1:

$$\tau_x^w = \left| \frac{\sum_{i<j} A_{ij} \left(b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)} \right) w_i}{\max\left(d_K^w\right)} \right| \leq 1.$$

Taking into account that $\max(d_K^w) = 2\sum_{i=1}^{m-1}(m-i)w_i = 2\sum_{i<j} w_i$, let us demonstrate that:

$$\tau_x^w = \left| \frac{\sum_{i<j} A_{ij} \left(b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)} \right) w_i}{2\sum_{i<j} w_i} \right| \leq 1.$$

Proof

- τ_x^w reaches the maximum value, equal to 1, if and only if the judges agree on all items. Therefore, $a_{ij}^{(\sigma_2)} = b_{ij}^{(\sigma_1)} = A_{ij} \quad \forall i < j$:

$$\tau_x^w = \frac{\sum_{i<j}(1+1)w_i}{2\sum_{i<j} w_i} = 1.$$

- τ_x^w reaches the minimum value, equal to -1, if and only if the judges disagree on all items. Therefore, $a_{ij}^{(\sigma_2)} = b_{ij}^{(\sigma_1)} = -A_{ij} \quad \forall i < j$:

$$\tau_x^w = \frac{\sum_{i<j}(-1-1)w_i}{2\sum_{i<j} w_i} = -1.$$

- τ_x^w takes a value between -1 and 1, when the judges agree on some items and disagree on the others. A particular case is when they agree on half of the items and disagree on the other half, i.e.: $a_{ij}^{(\sigma_2)} = -b_{ij}^{(\sigma_1)} \quad \forall i < j$:

$$\tau_x^w = \frac{\sum_{i<j}(\pm 1 \mp 1)w_i}{2\sum_{i<j} w_i} = 0.$$

3.4 Correspondence between weighted and unweighted measures

When an equal importance is assigned to the items' position, i.e. $w_i = \frac{1}{(m-1)} \; \forall i = 1, 2, \ldots, m-1$, the weighted rank correlation coefficient is

equivalent to the rank correlation coefficient defined by Emond and Mason.

Proof. Thanks to the symmetry of the combined input matrix, τ_x can be expressed also as:

$$\tau_x = \sum_{\substack{i,j=1 \\ i \neq j}}^{m} \frac{a_{ij} b_{ij}}{m(m-1)} = 2 \sum_{i<j}^{m} \frac{a_{ij} b_{ij}}{m(m-1)},$$

$$\tau_x^w = \frac{\sum_{i<j}^{m} w_i a_{ij}^{(\sigma_1)} b_{ij}^{(\sigma_1)} + \sum_{i<j}^{m} w_i a_{ij}^{(\sigma_2)} b_{ij}^{(\sigma_2)}}{\max(d_k^w)}$$

$$= \frac{\sum_{i<j}^{m} w_i A_{ij} \left(b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)} \right)}{m} = \frac{1}{m(m-1)} \sum_{i<j}^{m} \left(b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)} \right) A_{ij}.$$

The term A_{ij} can be omitted since, for $i < j$, $A_{ij} = +1 \; \forall i, j$.
Hence, we have to assess if the following equality holds:

$$2 \sum_{i<j}^{m} \frac{a_{ij} b_{ij}}{m(m-1)} = \frac{1}{m(m-1)} \sum_{i<j}^{m} \left(b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)} \right).$$

Fixing i and j, it sufficies to see what happens to each element of the equality:

$$2 a_{ij} b_{ij} = b_{ij}^{(\sigma_1)} + a_{ij}^{(\sigma_2)}.$$

Case 1. Both A and B prefer object i to j or vice versa. If $a_{ij} = b_{ij} = \pm 1$, it's obvious that $a_{ij}^{(\sigma_2)} = b_{ij}^{(\sigma_1)} = 1$ and substituting these values to the previous equation leads to obtain, respectively:

$$2 \cdot (\pm 1) \cdot (\pm 1) = 1 + 1.$$

Case 2. A prefers object i to j and B prefers j to i or vice versa. If $a_{ij} = \pm 1$ and $b_{ij} = \mp 1$, it's obvious that $a_{ij}^{(\sigma_2)} = b_{ij}^{(\sigma_1)} = -1$ and substituting these values to the previous equation leads to obtain, respectively:

$$2 \cdot (\pm 1) \cdot (\mp 1) = -1 - 1.$$

For equal weights assigned to the items ($w_i = \frac{1}{m-1}$, for each $i = 1, 2, \ldots, m-1$) the weighted distance is proportional to the Kemeny distance without weights, according to the number of items:

$$d_K^w = \frac{d_K}{m-1}.$$

Proof. Let's consider the natural increasing ordering of m elements $a = \{1, 2, \ldots, m\}$ and any other permutation b. The couples that contribute to the calculation of the Kemeny distance are only the discordant ones of b and they could be at most $\binom{m}{2}$. If a and b present $\binom{k}{2}$ discordant couples, with $(k = 2, 3, \ldots, m)$, i.e. if $b = \{m, m-1, m-2, \ldots, 2, 1\}$, we obtain:

$$d_k = \sum_{i<j}^{m} |a_{ij} - b_{ij}| = 2 \cdot \binom{k}{2},$$

$$d_k^w = \frac{1}{2} \sum_{i<j}^{m} w_i \left[|A_{ij} - b_{ij}^{(\sigma_1)}| + |A_{ij} - a_{ij}^{(\sigma_2)}| \right] =$$

$$= \frac{1}{2} \frac{1}{m-1} \sum_{i<j}^{m} \left[|A_{ij} - b_{ij}^{(\sigma_1)}| + |A_{ij} - a_{ij}^{(\sigma_2)}| \right] =$$

$$= \frac{1}{2} \cdot \frac{1}{m-1} [(2+2)\binom{k}{2}] = \frac{2}{m-1}\binom{k}{2} = \frac{d_k}{m-1}.$$

The previous results lead to the following considerations:

$$\max(d_K) = 2 \cdot \binom{m}{2},$$

$$\max(d_K^w) = \frac{\max(d_K)}{m-1}.$$

3.5 The consensus ranking problem and a suitable branch-and-bound BB algorithm

When dealing with preference rankings, searching for a ranking representative of a group of judges is a central theme (D'ambrosio *et al.*, 2015a). Two broad classes of approaches to consensus can be found in the literature: ad hoc procedures developed over time (parliamentary procedures, preferential voting needs, and so on) and more formal methodologies, based on a measure of distance (Cook, 2006). Here we follow this second approach, and, among the several consensus ranking measures proposed in the literature (Kemeny and Snell, 1962), the *median ranking* approach

will be used in the examples presented in the next subsection. It is defined as the ranking corresponding to the minimum sum of the weighted distances of all rankings from it.
Emond and Mason (2002) proposed to use a branch and bound algorithm, in order to avoid the research of the median ranking among all possible rankings belonging to Z^m or S^m (see Section 3). Hence, to compute the median ranking we start from the above cited branch and bound algorithm, implemented in the R package "ConsRank" (D'Ambrosio *et al.*, 2015b).
The proposed weighted correlation coefficient can be used to deal with a consensus ranking problem: given a $n \times m$ matrix $\mathbf{X}$, whose l-th row represents the ranking associated to the l-th judge, the purpose is to identify that ranking (b) (a candidate within the universe of the permutations of m elements) that best represents the average consensus of the subjects involved (i.e. the matrix $\mathbf{X}$). Considering that there is a one-to-one correspondence between a rank correlation coefficient and a distance, the solution ranking is reached by minimizing the average distance or, similarly, maximizing the average rank correlation:

$$\begin{aligned} \sum_{l=1}^{n} d({}_l x, b) &= \min, \\ \sum_{l=1}^{n} \tau_x^w({}_l x, b) &= \max . \end{aligned} \tag{3.4}$$

Maximizing the expression in Equation (3.4) means to maximize the following function, once a candidate ranking is fixed among all $m!$ permutations of m items:

$$\sum_{l=1}^{n} \tau_x^w({}_l x, b) = \sum_{l=1}^{n} \frac{\sum_{i<j}^{m} a_{ij}^{X(l)}(l) b_{ij}^{X(l)} w_i + \sum_{i<j}^{m} a_{ij}^{b}(l) b_{ij}^{b} w_i}{\max[d_K^w(\mathbf{X}, b)]},$$

where l refers to one of the n orderings, $a_{ij}^{X(l)}(l)$ and b_{ij}^{b} are the ij-th elements of the score matrices related to the natural ascending orderings $(1, 2, \ldots, m)$, named A_{ij}, $b_{ij}^{X(l)}$ is the generic ij-th element in the score matrix of the candidate ranking b reordered according to the l-th ordering of $\mathbf{X}$ and $a_{ij}^{b}(l)$ is the generic ij-th element in the score matrix of the l-th ranking in $\mathbf{X}$ reordered according to b, and the denominator is defined in Equation (3.6). Changing the order of the summation operations, leads

to work under another perspective:

$$\sum_{l=1}^{n} \tau_x^w({}_l x, b) =$$

$$= \frac{1}{\max\left[d_K^w(\mathbf{X}, b)\right]} \left[\sum_{i<j}^{m} \sum_{l=1}^{n} a_{ij}^{X(l)}(l) b_{ij}^{X(l)} w_i + \sum_{i<j}^{m} \sum_{l=1}^{n} a_{ij}^{b}(l) b_{ij}^{b} w_i \right] =$$

$$= \frac{1}{\max\left[d_K^w(\mathbf{X}, b)\right]} \left[\sum_{i<j}^{m} \sum_{l=1}^{n} b_{ij}^{X(l)} A_{ij} w_i + \sum_{i<j}^{m} \sum_{l=1}^{n} a_{ij}^{b}(l) A_{ij} w_i \right] =$$

$$= \frac{1}{\max\left[d_K^w(\mathbf{X}, b)\right]} \left[\sum_{i<j}^{m} c_{ij}(b) A_{ij} w_i + \sum_{i<j}^{m} c_{ij}(a) A_{ij} w_i \right], \qquad (3.5)$$

where $c_{ij}(b) = \sum_{l=1}^{n} b_{ij}^{X(l)}$ and $c_{ij}(a) = \sum_{l=1}^{n} a_{ij}^{b}(l)$ are the elements of the combined input matrices (Emond and Mason, 2002). In few words, they are $m \times m$ matrices obtained aggregating the score matrices of all the individual orderings. Hence, we deal with the optimization of a function given by two components, where each one is a product of combine input matrices, score matrices (A_{ij}) and the vector of weights. The maximization problem is limited to maximizing only the numerator, because the denominator is a fixed quantity depending on the number of subjects, the number of items and the positional weights fixed at the beginning of the process.
Emond and Mason (2002) proposed the BB algorithm to deal with the *consensus ranking problem*. Recently, Amodio *et al.* (2016) and D'ambrosio *et al.* (2015a) proposed two accurate algorithms, they called QUICK and FAST, for identifying the median ranking when dealing with weak and partial rankings, in the framework of the Kemeny approach.

The procedure proposed here is based on their approach, but τ_x is replaced with τ_x^w; in particular, following Emond and Mason (2002), we calculate the maximum possible value P^* of the numerator in Equation (3.5), which is represented by the denominator since a rank correlation coefficient has a maximum value of $|\pm 1|$:

$$P^* = \max\left[d_K^w(\mathbf{X}, b)\right] = 2n \sum_{i=1}^{m-1} (m - i) w_i, \qquad (3.6)$$

and taking a candidate ranking among the $m!$ permutations of the items, we evaluate the numerator in the Equation (3.5):

$$p = \sum_{i<j}^{m} c_{ij}(b) A_{ij} w_i + \sum_{i<j}^{m} c_{ij}(a) A_{ij} w_i = \sum_{i<j}^{m} c_{ij}^{w}. \tag{3.7}$$

In particular, to identify a good candidate to be the median ranking that can be used as an input in the algorithm, we follow (Amodio *et al.*, 2016).

The two quantities (P^* and p) are used for measuring an initial penalization, in the consensus ranking process, in the following way:

$$P = P^* - p. \tag{3.8}$$

The purpose of the algorithm is to find, among all the possible linear rankings, the one that provides the minimum penalty from now on. In order to achieve this objective, the set of $m!$ permutations is divided into two mutually exclusive branches according to the position of the two first items (namely i and j) in the ordering used as the initial solution. After that, an incremental penalty for each branch can be computed, taking into account the c_{ij}^{w} and c_{ji}^{w} of the entire C^w input matrix (defined in Equation (3.7)):

BRANCH 1

- object i is preferred to object j:

  a) if $c_{ij}^{w} > 0$ and $c_{ji}^{w} > 0$, with $c_{ij}^{w} \geq c_{ji}^{w}$, then $\delta P = 0$

  b) if $c_{ij}^{w} < 0$ and $c_{ji}^{w} < 0$, then $\delta P = |c_{ij}^{w} + c_{ji}^{w}|$

BRANCH 2

- object j is preferred to object i:

  a) if $c_{ij}^{w} < 0$ and $c_{ji}^{w} < 0$, with $c_{ij}^{w} \leq c_{ji}^{w}$, then $\delta P = |c_{ij}^{w} + c_{ji}^{w}|$

  b) if $c_{ij}^{w} > 0$ and $c_{ji}^{w} > 0$, then $\delta P = 0$

where δP represents the incremental penalty. If the incremental penalty obtained from a branch is greater than the initial penalty, then the rankings belonging to that branch are excluded from the search process. In the opposite case, if a branch provides a smaller (or equal) penalty than the initial one, then the next object in the initial solution is taken into

consideration and other new branches are built up, by moving this object in all possible positions with respect to the objects considered before. The incremental penalty produced by the new branches is again the tool for cutting useless branches and for keeping, only, the useful ones until a solution is reached. Of course, with constant weights ($w_i = \frac{1}{m-1}$) τ_x^w reduces to τ_x (see Proposition 2 in Section 3), therefore the modified BB algorithm proposed loses utility.

3.6 Experimental evaluation

To assess the performance of our algorithm, we considered both a simulation study and real datasets. We implemented the proposed BB algorithm in R environments by suitably modifying the corresponding functions of the "ConsRank" package (D'Ambrosio *et al.*, 2015b).

Simulations

In the simulation study ranking data were generated according to a Mallows model (Irurozki *et al.*, 2016), which is an exponential model defined by a central permutation α and a spread (or dispersion) parameter θ. When $\theta > 0$ (with $\theta = 0$ we obtain the uniform distribution), α is the mode of the distribution, i.e., the permutation with the highest probability. The probability of any other permutation S decays exponentially as its distance to the central permutation increases, according to the model

$$f_\theta(\alpha; S) = C(\theta) \exp(-\theta d(S, \alpha)), \tag{3.9}$$

where $C(\theta)$ is a constant of normalization. The distance d can be measured in many ways, including Kendall's, Cayley, Hamming and Ulam distances. The Mallows model that uses Kendall's distance (used in this chapter) is also known in the literature as the Mallows ϕ model.
The four levels chosen for θ were 0.4, 0.7, 1 and 2. Two different levels for the number of items were taken into account: 5 and 9. The position weighting vectors were employed according to a precise structure, let's show those relative to 5 items: $w_1 = (1/4, 1/4, 1/4, 1/4)$, $w_2 = (4/10, 3/10, 2/10, 1/10)$, $w_3 = (1/2, 1/2, 0, 0)$, $w_4 = (2/3, 1/3, 0, 0)$ and $w_5 = (1, 0, 0, 0)$. In other words, equal and decreasing weights were considered, at first involving all the weights and then only half of the total amount of the vectors, and, finally, only the first position was weighted. The sample size used for all the datasets generated was 50, and for each combination of θ and the number of items, 100 samples were generated.

For each sample, we estimated the consensus ranking and the corresponding τ_x^w (8) with each weighting vector. Figure 3.1 compares the true distribution (i.e., computed with reference to the true mode α used to generate the data according to model (3.9)) of τ_x^w, always shown in white colour, and τ_x^w computed with respect to the estimated consensus; on the left rankings of 5 items are considered, 9 items on the right.

As it appears, the implemented procedure always finds the correct consensus, since the distributions of true and estimated τ_x^w are comparable. Both with 5 and with 9 items the higher θ the higher τ_x^w and the simpler the weighting vector (few positions involved) the higher τ_x^w.

It is possible (see for example the lowest values Figure 3.1-left, $\theta = 0.4$ or $\theta = 0.7$ and $w = w_2$ to $w = w_5$) that the estimated consensus is even better (i.e. with a higher τ_x^w) than the true model: this is due to the introduction of weights with a simpler structure.

Real data

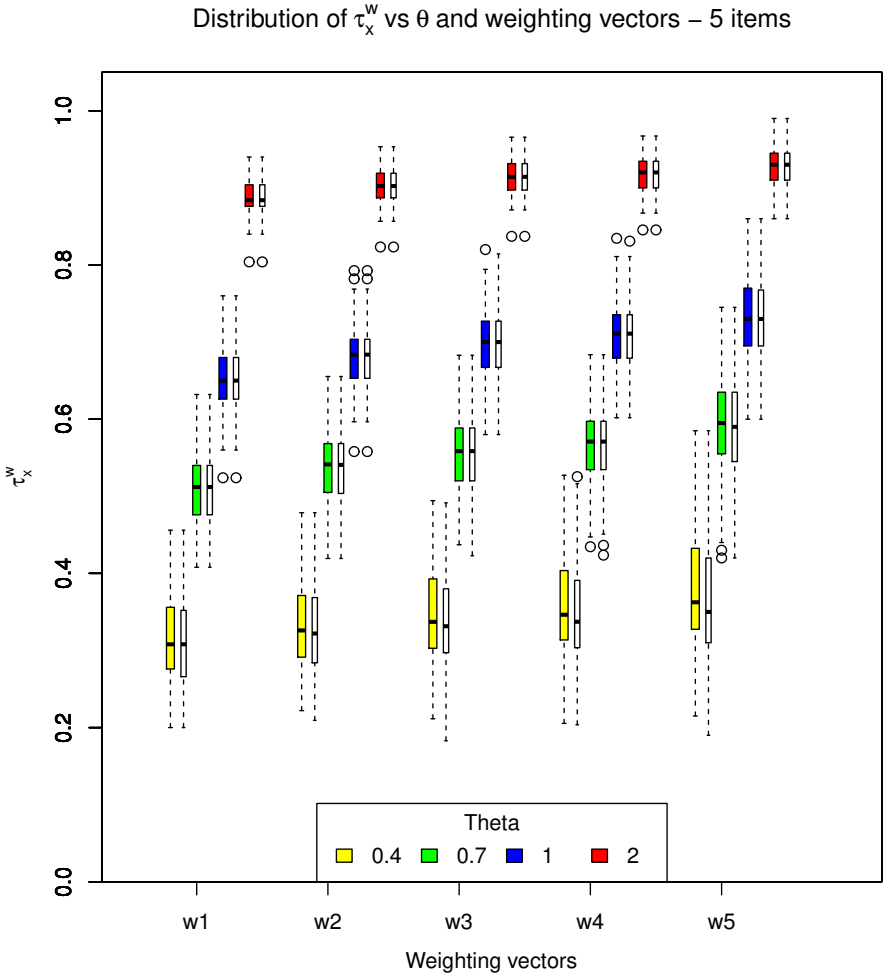
The first dataset considered is AGH Course Selection (*http://www.preflib.org/data/election/agh/*). This dataset contains the results of a survey conducted among students at AGU University of Science and Technology regarding their course preferences. Each student provided a rank ordering over all the courses with no missing elements. In the dataset of 2003, here considered, there are 9 courses (coded C1 to C9) to choose from, i.e each student had to order 9 courses according to his/her preference.

As previously considered in the simulations, 5 different weighting vectors are considered (shown in Table 3.1), in order to assess their effect on the estimated consensus: with w_1 we give the same importance to each position, which means that we do not weigh positions (see Proposition 3); with w_5 we give importance to the first position only.

w_1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
w_2	0.222	0.194	0.167	0.139	0.111	0.083	0.056	0.028
w_3	0.250	0.250	0.250	0.250	0.000	0.000	0.000	0.000
w_4	0.400	0.300	0.200	0.100	0.000	0.000	0.000	0.000
w_5	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 3.1: AGH Course Selection dataset: Weighting vectors

The consensus estimated for each weighting vector is shown in Table 3.2. With the first four weighting vectors we obtain very similar results, with increasing values of τ_x^w (as expected according to the simulation


Distribution of τ_x^w vs θ and weighting vectors – 5 items
τ_x^w
0.0
0.2
0.4
0.6
0.8
1.0
Theta
0.4
0.7
1
2
w1
w2
w3
w4
w5
Weighting vectors

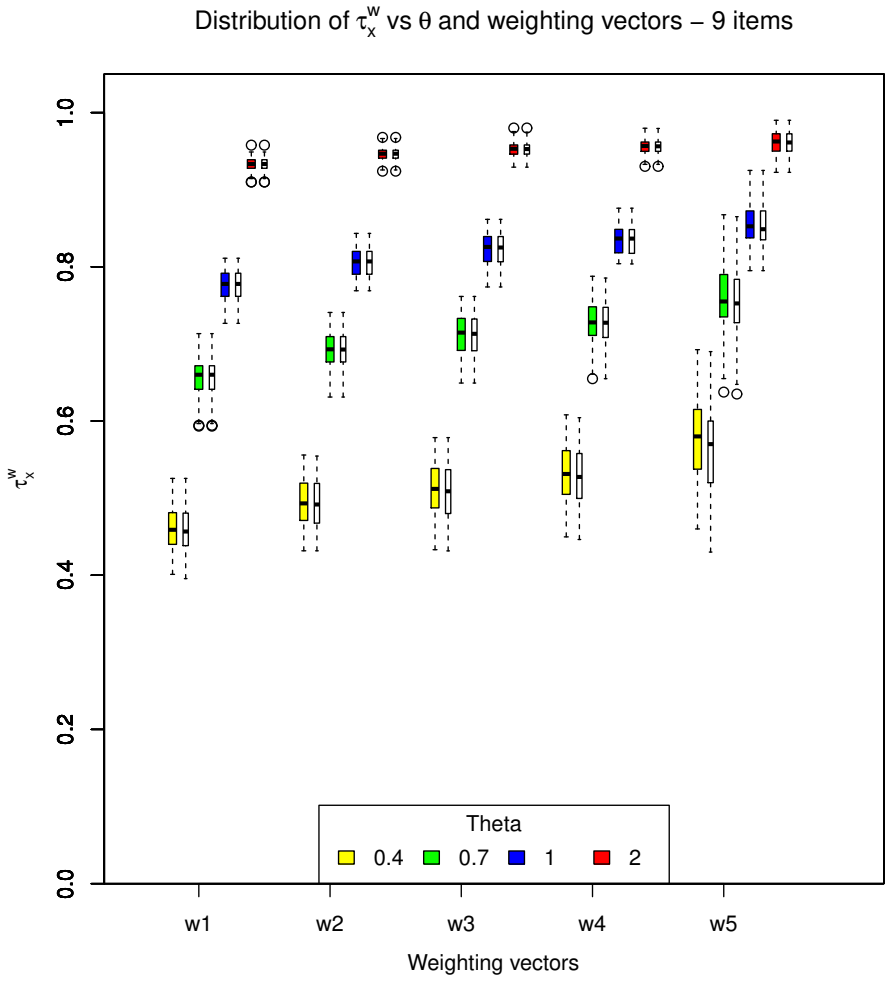


FIGURE 3.1: Real (white color) and estimated tau_x^w distribution vs $theta$ and weighting vectors for rankings of 5 items (left) and 9 items (right).

results). With w_5 we obtain 16 different solutions corresponding to the same value of τ_x^w, always with course "C1" in first position.

	1	2	3	4	5	6	7	8	9	τ_x^w
w_1	C1	C9	C2	C3	C7	C8	C4	C6	C5	0.571
w_2	C1	C9	C2	C3	C7	C8	C4	C6	C5	0.603
w_3	C1	C9	C2	C4	C7	C8	C5	C6	C3	0.606
w_4	C1	C9	C2	C6	C7	C8	C4	C5	C3	0.680
w_5	C1	C6	C2	C9	C7	C8	C4	C3	C5	0.955
	C1	C6	C2	C8	C9	C7	C3	C4	C5	
	C1	C7	C2	C8	C6	C9	C3	C4	C5	
	C1	C9	C2	C7	C8	C6	C3	C4	C5	
	C1	C8	C2	C9	C6	C7	C3	C4	C5	
	C1	C9	C2	C6	C7	C8	C3	C4	C5	
	C1	C7	C2	C9	C8	C6	C3	C4	C5	
	C1	C9	C2	C7	C8	C6	C4	C3	C5	
	C1	C8	C2	C9	C7	C6	C4	C3	C5	
	C1	C6	C2	C7	C9	C8	C3	C4	C5	
	C1	C7	C2	C6	C8	C9	C4	C3	C5	
	C1	C7	C2	C6	C8	C9	C3	C4	C5	
	C1	C7	C2	C6	C9	C8	C4	C3	C5	
	C1	C6	C2	C7	C8	C9	C3	C4	C5	
	C1	C6	C2	C9	C7	C8	C3	C4	C5	
	C1	C8	C2	C6	C9	C7	C4	C3	C5	

TABLE 3.2: Consensus rankings for each weighting vectors

The second dataset considered is the T-shirt dataset (*http://www.preflib.org/data/election/shirt/*) that contains complete rank orderings of T-Shirt designs voted on by members of the Optimization Research Group at NICTA. There are 11 designs (candidates, coded A to K) and 30 votes about these designs. Voters were required to submit complete strict orders.

Again 5 different weighting vectors are considered (shown in Table 3.3), in order to assess their effect on the estimated consensus.: with w_1 we give the same importance to each position, which means that we do not weigh positions (see Proposition 3); with w_5 we give importance to the first position only.

This time the solutions are quite different (Table 3.4): while t-shirt coded E is in first position with weight w_1 (or analogously without

w_1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125
w_2	0.222	0.194	0.167	0.139	0.111	0.083	0.056	0.028
w_3	0.250	0.250	0.250	0.250	0.000	0.000	0.000	0.000
w_4	0.400	0.300	0.200	0.100	0.000	0.000	0.000	0.000
w_5	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

TABLE 3.3: T-shirt dataset: eighting vectors

	1	2	3	4	5	6	7	8	9	10	11	tau
w_1	E	F	L	I	J	K	C	A	D	B	H	0.189
	E	F	L	I	J	K	B	A	D	C	H	
w_2	C	D	L	I	J	K	B	A	F	E	H	0.2
w_3	C	D	K	H	J	L	B	A	F	E	I	0.201
	C	D	L	H	J	K	B	A	F	E	I	
	C	D	L	I	J	K	B	A	F	E	H	
w_4	C	D	L	H	J	K	B	A	F	E	I	0.211
w_5	B	C	D	E	L	H	J	A	F	I	K	0.343
	B	C	E	D	K	H	I	A	F	J	L	
	B	C	D	E	K	H	J	A	F	I	L	
	B	C	E	D	L	H	I	A	F	J	K	

TABLE 3.4: T shirt results

weights), t-shirt C becomes the most preferred with weights w_2, w_3 and w_4, and t-shirt B is in first position with weight w_5.

3.7 Conclusion

We have proposed a new position weighed rank correlation coefficient for linear orders. We demonstrated that the proposed coefficient is in one to one correspondence with the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010), whereby if equal importance is assigned to an items' position, the weighted rank correlation coefficient is equivalent to the rank correlation coefficient defined by Emond and Mason (2002).

The proposed weighted correlation coefficient can be used to deal with a consensus ranking problem: given n linear orderings of m objects, the purpose is to identify the ranking (b) that best represents the average consensus of the subjects involved.

We implemented the proposed BB algorithm in R environments,by

suitably modifying the functions of the "ConsRank" package (D'Ambrosio *et al.*, 2015b). A simulation plan shows that the implemented procedure always find the correct consensus, since the distributions of true and estimated τ_x^w are comparable. Both with 5 and with 9 items the higher θ (the dispersion parameter, indicating the level of heterogeneity/homogeneity among rankings) the higher τ_x^w and the simpler the weighting vector (few positions involved) the higher τ_x^w. Finally, applying the proposed procedure to two real datasets, we have shown how a different weighting vector can modify the estimated consensus.

Chapter 4

A position weighted rank coefficient for rankings with ties

4.1 Weak orders

Combining the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010) and the extension of τ_x provided by Emond and Mason (2002), we propose a new rank correlation coefficient working with a couple of score matrices.

Let's define the generic (i, j) element of the score matrices a_{ij}^+ and a_{ij}^- related to a ranking a as follows:

$$a_{ij}^+, b_{ij}^+ = \begin{cases} 1 & i \text{ preferred to or tied with } j \\ 0 & i = j \\ -1 & j \text{ preferred to } i \end{cases} \qquad a_{ij}^-, b_{ij}^- = \begin{cases} 1 & i \text{ preferred to } j \\ 0 & i = j \\ -1 & j \text{ preferred to or tied with } i \end{cases}$$

Following the observations in Emond and Mason (2002) (Sections 38, 39), both "+1" in a_{ij}^+ and "-1" in a_{ij}^- are associated to ties. The new rank correlation coefficient uses both these score matrices and it is defined as:

$$\tau_x^w(a, b) = \frac{\sum_{i<j}^m \left(a_{ij}^{+\sigma_1} b_{ij}^{+\sigma_1} + a_{ij}^{+\sigma_2} b_{ij}^{+\sigma_2} + a_{ij}^{-\sigma_1} b_{ij}^{-\sigma_1} + a_{ij}^{-\sigma_2} b_{ij}^{-\sigma_2} \right) w_i}{2 \max \left[d_K^w \right]}, \tag{4.1}$$

where the denominator represents twice the maximum value of the Kemeny weighted distances (see Equation (3.2)).

A question that arises immediately is: what weight or weights are used when two items occupy the same position in a ranking? Equation (4.1), based on the two newly introduced score matrices, associates the same weight, as the next example shows. Let's consider $b = 1, 2, 3, 4$, and

$a_1 = 1,3,2,4$, $a_2 = 1,2,2,3$, $a_3 = 1,1,2,3$ and $w = (w_1, 0, 0)$. Giving weight to the first position only, the correlation between a_1 and b, as well as the one between a_2 and b has to be maximum, while the correlation between a_3 and b has not. And sure enough it results to be:

$\tau_x^w(a_1, b) = \tau_x^w(a_2, b) = 1$, while, $\tau_x^w(a_3, b) < 1$.

4.2 Correspondence between distance and correlation

We will demonstrate that Equation (4.1) is the correlation coefficient corresponding to the weighted Kemeny distance in Equation (2.10) through the straightforward linear transformation:

$$\frac{\sum_{i<j}^{m}(a_{ij}^{+\sigma_1}b_{ij}^{+\sigma_1} + a_{ij}^{+\sigma_2}b_{ij}^{+\sigma_2} + a_{ij}^{-\sigma_1}b_{ij}^{-\sigma_1} + a_{ij}^{-\sigma_2}b_{ij}^{-\sigma_2})w_i}{2\max\left[d_K^w\right]} = 1 - \frac{2d_K^w}{\max\left[d_K^w\right]},$$

or equivalently

$$\sum_{i<j}^{m}(a_{ij}^{+\sigma_1}b_{ij}^{+\sigma_1} + a_{ij}^{+\sigma_2}b_{ij}^{+\sigma_2} + a_{ij}^{-\sigma_1}b_{ij}^{-\sigma_1} + a_{ij}^{-\sigma_2}b_{ij}^{-\sigma_2})w_i = 2\max\left[d_K^w\right] - 4d_K^w, \tag{4.2}$$

where $\max\left[d_K^w\right]$ and d_K^w are defined in Equation (3.2) and in Equation (2.10), respectively.

According to Emond and Mason (2002), if two rankings a and b agree except for a set S of k objects, which is a segment of both, then $d_K^w(a, b)$ may be computed as if these k objects were the only objects being ranked. As a consequence, to prove the equality in (4.2) we will show that for each pair of objects i and j:

$$a_{ij}^{+\sigma_1}b_{ij}^{+\sigma_1} + a_{ij}^{+\sigma_2}b_{ij}^{+\sigma_2} + a_{ij}^{-\sigma_1}b_{ij}^{-\sigma_1} + a_{ij}^{-\sigma_2}b_{ij}^{-\sigma_2} = 4(m-i) - 2\left[|a_{ij}^{\sigma_1} - b_{ij}^{\sigma_1}| + |b_{ij}^{\sigma_2} - a_{ij}^{\sigma_2}|\right] \tag{4.3}$$

In Equation (4.3) the weights w_i have been omitted from both the sides. There are nine possible combinations of orderings for item i and j between judges ***a*** and ***b***, but only four distinct cases must be considered. The other five are equivalent to one of these four through a simple relabeling of the rankers and/or the objects. (Emond and Mason, 2002).

Case 1. Both ***a*** and ***b*** prefer object i to j. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_2} = 1$. The τ_x^w score matrix values are: $a_{ij}^{+\sigma_1} = b_{ij}^{+\sigma_1} = a_{ij}^{+\sigma_2} = b_{ij}^{+\sigma_2} = a_{ij}^{-\sigma_1} = b_{ij}^{-\sigma_1} = a_{ij}^{-\sigma_2} = b_{ij}^{-\sigma_2} = 1$. Hence, the equality in Equation (4.3) holds:

$1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 1 = 4 - 2[|1 - (1)| + |1 - (1)|]$.

Case 2. ***a*** prefers object i to j and ***b*** prefers the two objects as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = a_{ij}^{\sigma_2} = 1$ and $b_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 0$. The τ_x^w score matrix values are: $a_{ij}^{+\sigma_1} = b_{ij}^{+\sigma_1} = a_{ij}^{+\sigma_2} = b_{ij}^{+\sigma_2} = a_{ij}^{-\sigma_1} = a_{ij}^{-\sigma_2} = 1$ and $b_{ij}^{-\sigma_1} = b_{ij}^{-\sigma_2} = -1$. Hence, the equality in Equation (4.3) holds:
$1 \cdot 1 + 1 \cdot 1 + 1 \cdot (-1) + 1 \cdot (-1) = 4 - 2[|1 - 0| + |1 - 0|]$.

Case 3. ***a*** prefers object i to j and ***b*** prefers j to object i. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = 1$ and $a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = -1$. The τ_x^w score matrix values are: $a_{ij}^{+\sigma_1} = b_{ij}^{+\sigma_2} = a_{ij}^{-\sigma_1} = b_{ij}^{-\sigma_2} = 1$ and $a_{ij}^{+\sigma_2} = b_{ij}^{+\sigma_2} = a_{ij}^{-\sigma_2} = b_{ij}^{-\sigma_1} = -1$. Hence, the equality in Equation (4.3) holds:
$1 \cdot (-1) + (-1) \cdot 1 + 1 \cdot (-1) + (-1) \cdot (1) = 4 - 2[|1 - (-1)| + |1 - (-1)|]$.

Case 4. Both judges ***a*** and ***b*** rank the objects i and j as tied. The Kemeny-Snell matrix values are: $a_{ij}^{\sigma_1} = b_{ij}^{\sigma_2} = a_{ij}^{\sigma_2} = b_{ij}^{\sigma_1} = 0$. The τ_x^w score matrix values are: $a_{ij}^{+\sigma_1} = b_{ij}^{+\sigma_1} = a_{ij}^{+\sigma_2} = b_{ij}^{+\sigma_2} = 1$ and $a_{ij}^{-\sigma_1} = b_{ij}^{-\sigma_1} = a_{ij}^{-\sigma_2} = b_{ij}^{-\sigma_2} = -1$. Hence, the equality in Equation (4.3) holds:
$1 \cdot (1) + (1) \cdot 1 + 1 \cdot (1) + (1) \cdot (1) = 4 - 2[|0 - 0| + |0 - 0|]$.

4.3 Minimum and Maximum values

From the demonstrations in Section 4.2, τ_x^w assumes its maximum value, equal to 1, if and only if for all i and j only *Case 1* or only *Case 4* are observed. Therefore, differently from what happens with Kendall τ_b (see (Emond and Mason, 2002) sect 3.3), τ_x^w assumes the maximum value even when a generic all tied ranking is compared with itself.

Analogously, τ_x^w can be minimum, and equal to -1, if and only if for all i and j only *Case 3* occurs.

4.4 Correspondence between weighted and unweighted measures

For equal weights assigned to the items ($w_i = \frac{1}{m-1}$, for each $i = 1, 2, \ldots, m-1$) the weighted distance is proportional to the classical Kemeny distance, on the basis of the number of items:

$$d_x^w = \frac{d_x}{m-1}$$

Proof. Referring to the cases listed in Section 4.2:

Case 1. $d_x^w = \frac{1}{2}[|1-(1)|+|1-(1)|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Case 2. $d_x^w = \frac{1}{2}[|1-0|+|1-0|]w_i = \frac{1}{m-1}$ and $d_x = \frac{1}{2}[|1-0|+|1-0|] = 1$

Case 3. $d_x^w = \frac{1}{2}[|1-(-1)|+|1-(-1)|]w_i = \frac{2}{m-1}$ and $d_x = \frac{1}{2}[|1-(-1)|+|1-(-1)|] = 2$

Case 4. $d_x^w = \frac{1}{2}[|0-0|+|0-0|]w_i = 0$ and $d_x = \frac{1}{2}[|0-0|+|0-0|] = 0$

Corollary Since $\tau_x \leftrightarrow d_K$ and $\tau_x^w \leftrightarrow d_K^w$, then the weighted rank correlation coefficient is equivalent to the rank correlation coefficient defined by Emond and Mason when equal importance is given to the positions occupied by the items:

$$\tau_x^w = \tau_x, \quad \text{with } w_i = \frac{1}{m-1} \quad \forall i = 1, 2, \ldots, m-1$$

4.5 The consensus ranking problem and a suitable branch-and-bound BB algorithm

The proposed weighted correlation coefficient can be used to deal with a consensus ranking problem: given n rankings, full or weak, of m items, which best represents the consensus opinion? This consensus will be the ranking that shows the maximum correlation, with the whole set of n rankings.

Given a $n \times m$ matrix $\mathbf{X}$, whose l-th row represents the ranking associated to the l-th judge, the consensus ranking, i.e. the ranking r (a candidate within the universe of the permutations with ties of m elements) that best represents the matrix $\mathbf{X}$, is that ranking that maximizes the following expression:

$$\max \sum_{l=1}^{n} \tau_x^w \left({}_l x, r\right) = \tag{4.4}$$

$$= \max \sum_{l=1}^{n} \frac{\sum_{i<j}^{m} \left({}_l x_{ij}^{+\sigma_l} r_{ij}^{+\sigma_l} +_l x_{ij}^{+\sigma_r} r_{ij}^{+\sigma_r} +_l x_{ij}^{-\sigma_l} r_{ij}^{-\sigma_l} +_l x_{ij}^{-\sigma_r} r_{ij}^{-\sigma_r} \right) w_i}{2 \max \left[d_K^w(X, r) \right]}$$

where $\max[d_K^w(X,r)]$ is $2n \sum_{i=1}^{m-1}(m-i)w_i$.

Changing the order of the summations, leads to work under another perspective:

$$\sum_{l=1}^{n} \tau_x^w \left({}_l x, r\right) = \tag{4.5}$$

$$= \sum_{l=1}^{n} \frac{\sum_{i<j}^{m} \left({}_l x_{ij}^{+\sigma_l} r_{ij}^{+\sigma_l} +_l x_{ij}^{+\sigma_r} r_{ij}^{+\sigma_r} +_l x_{ij}^{-\sigma_l} r_{ij}^{-\sigma_l} +_l x_{ij}^{-\sigma_r} r_{ij}^{-\sigma_r} \right) w_i}{2 \max \left[d_K^w(X, r) \right]} =$$

$$= \frac{1}{2 \max \left[d_K^w(X, r) \right]} \cdot$$

$$\cdot \sum_{i<j}^{m} \left[\sum_{l=1}^{n} {}_l x_{ij}^{+\sigma_l} r_{ij}^{+\sigma_l} +_l x_{ij}^{+\sigma_r} r_{ij}^{+\sigma_r} +_l x_{ij}^{-\sigma_l} r_{ij}^{-\sigma_l} +_l x_{ij}^{-\sigma_r} r_{ij}^{-\sigma_r} w_i \right] =$$

$$= \frac{1}{2 \max \left[d_K^w(X, r) \right]} \left[\sum_{i<j}^{m} c_{ij}^w \right]$$

where c_{ij}^w is the analogous of a combined input matrix (Emond and Mason, 2002). In few words, it is $m \times m$ matrix obtained aggregating the score matrices of all the individual orderings.

The maximization problem reduces to maximize only the numerator, because the denominator is a fixed quantity depending on the number of subjects, the number of items and the positional weights fixed at the beginning of the process.

Emond and Mason (2002) proposed a BB algorithm to deal with the *consensus ranking problem*. As we mentioned before, Amodio *et al.* (2016) and D'Ambrosio *et al.* (2015b) proposed two accurate algorithms, they called QUICK and FAST, for identifying the median ranking when dealing with weak and partial rankings, in the framework of the kemeny approah.

The procedure proposed is based, again, on their approach, but τ_x is replaced with τ_x^w; in particular, following Emond and Mason (2002), the maximum possible value P^* of the numerator in Equation (4.5) is calculated; it is represented by the denominator of the equation itself,

since a rank correlation coefficient has a maximum value equal to $|\pm 1|$:

$$P^* = \max\left[d_K^w(X,b)\right] = 2n\sum_{i=1}^{m-1}(m-i)w_i,$$

and taking a candidate ranking within the universe of the permutations with ties of the items, we evaluate the numerator in the Equation (4.5):

$$p = \left[\sum_{i<j}^{m} c_{ij}^w\right].$$

The two quantities P^* and p are used for estabilish an initial penalization, in the consensus ranking process, in the following way:

$$P = P^* - p. \tag{4.6}$$

The purpose of the algorithm is to find out, among all the possible weak rankings, that one which provides the minimum penalty. In order to reach this objective, the universe of the permutations with ties is divided into three mutually exclusive branches according to the position of the two first items (named i and j) in the ordering used as the initial solution. After that, an incremental penalty for each branch can be computed, taking into account the c_{ij}^w and c_{ji}^w of the entire C^w input matrix:

BRANCH 1

- object i is preferred to object j:
  a) if $c_{ij}^w > c_{ji}^w > 0$, then $\delta P = 0$
  b) if $c_{ij}^w = c_{ji}^w = 0$, then $\delta P = [(w_i + w_j)\cdot(m-1)\cdot n]/2$
  c) if $c_{ij}^w < c_{ji}^w < 0$, then $\delta P = |c_{ij}^w + c_{ji}^w|$

BRANCH 2

- object i is tied to object j:
  a) if $c_{ij}^w = c_{ji}^w = 0$, then $\delta P = [(w_i + w_j)\cdot(m-1)\cdot n]/2$
  b) if $c_{ij}^w > 0$ and $c_{ji}^w < 0$, then $\delta P = 0$
  c) if $c_{ij}^w = c_{ji}^w = 0$, then $\delta P = [(w_i + w_j)\cdot(m-1)\cdot n]/2$

BRANCH 3

- object j is preferred to object i:

  a) if $c_{ij}^w < c_{ji}^w < 0$, then $\delta P = |c_{ij}^w + c_{ji}^w|$
  b) if $c_{ij}^w = c_{ji}^w = 0$, then $\delta P = [(w_i + w_j) \cdot (m-1) \cdot n]/2$
  c) if $c_{ij}^w > c_{ji}^w > 0$, then $\delta P = 0$

where δP represents the incremental penalty. If the incremental penalty computed from a branch is greater than the initial penalty, then the rankings belonging to that branch are excluded from the searching process. Otherwise, if a branch provides a smaller (or equal) penalty than the initial one, then the next object in the initial solution is considered and other new branches are built up, by moving this object in all possible positions with respect to the objects considered before. The incremental penalty produced by the new branches is again the tool for cutting useless branches and for holding, only, the useful ones until a solution is reached. Actually, following Amodio *et al.* (2016) QUICK algorithm, the penalty is evaluated by considering all items in the input ranking, while in the original BB formulation, the penalty was computed by considering only the elements of the combined input matrix associated with the processed objects, adding up these partial values.

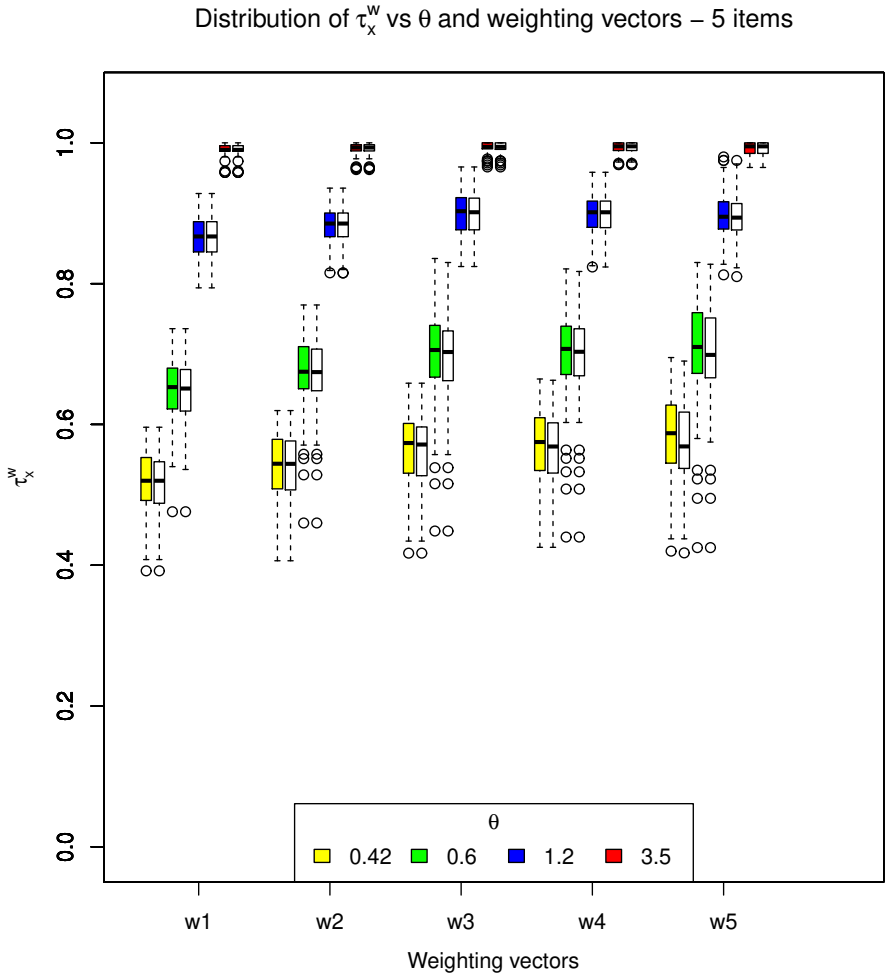
4.6 Experimental evaluation

Both a simulation study and real datasets will be considered to assess the performance of the proposed procedure. Data analysis is performed using our code written in R language. The proposed BB algorithm has been implemented in R environment by suitably modifying the corresponding functions of the "ConsRank" package (D'Ambrosio *et al.*, 2015b).

Simulations

In the simulation study, ranking data were generated according to the Mallows model (Irurozki *et al.*, 2016), which is an exponential model defined previosuly (see Equation (3.9), Section 3.6). We consider $m = 4$ and $m = 9$ items, while four different values of θ are used: 0.42, 0.6, 1.2 and 3.5. The full space of complete and tied rankings has been considered. The position weighting vectors were employed according to a precise structure: for example, for 5 items we consider $w_1 = (1/4, 1/4, 1/4, 1/4)$, $w_2 = (4/10, 3/10, 2/10, 1/10)$, $w_3 = (1/2, 1/2, 0, 0)$, $w_4 = (2/3, 1/3, 0, 0)$ and $w_5 = (1, 0, 0)$. In brief, equal and decreasing weightings were considered, at first involving all the weights and then only half of the total number of the positions

and, finally, only the first position was weighted. The sample size used for all the datasets generated was 50, and for each combination of θ and number of items, 100 samples were generated. For each sample, we estimated the consensus ranking and the corresponding τ_x^w in Equation (4.5) by using all the weighting vectors. Figure 4.1 compares the true distribution (i.e., computed with reference to the true mode α used to generate the data according to model (3.9)) of τ_x^w, always shown in white color, and τ_x^w computed with respect to the estimated consensus.



As it appears, the implemented procedure always finds the correct

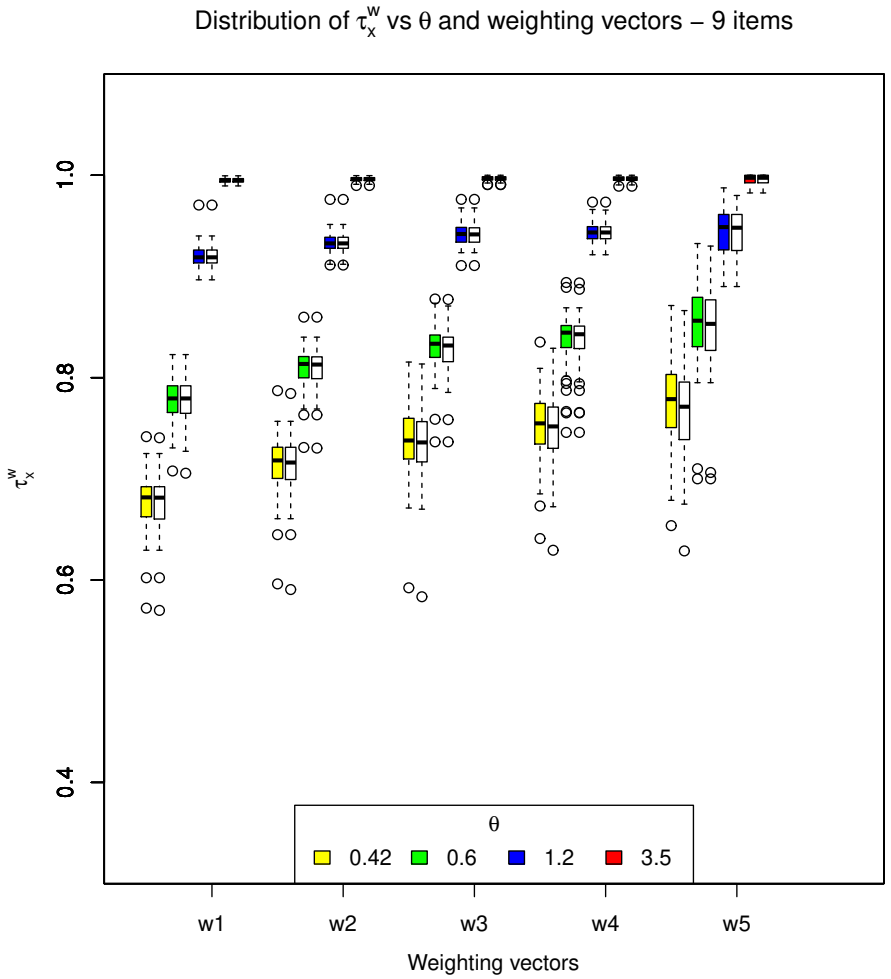


FIGURE 4.1: Real (white color) and estimated tau_x^w distribution vs $theta$ and weighting vectors for rankings of 5 items (left) and 9 items (right).

consensus, since the distributions of true and estimated τ_x^w are comparable. Both with 5 and with 9 items the higher θ the higher τ_x^w and the simpler the weighting vector (few positions involved) the higher τ_x^w.

Real data

The first dataset considered is the Sport Dataset (Amodio *et al.*, 2016). In this data, 130 students of the University of Illinois were asked to rank seven sports according to their preference for participate on. The sports considered were: A = baseball, B = football, C = basketball, D = tennis, E = cycling, F = swimming and G = jogging.

As previously considered in the simulations, 5 different weighting vectors are considered (shown in Table 4.1), in order to assess their effect on the estimated consensus.: with w_1 we give the same importance to each position, which means that we do not weight positions (see Section 4.4); with w_5 we give importance to the first position only.

w_1	0.125	0.125	0.125	0.125	0.125	0.125
w_2	0.286	0.238	0.190	0.143	0.095	0.048
w_3	0.334	0.333	0.333	0.000	0.000	0.000
w_4	0.500	0.266	0.133	0.000	0.000	0.000
w_5	1.000	0.000	0.000	0.000	0.000	0.000

TABLE 4.1: Sport dataset: Weighting vectors

w_1	$EFCADBG$
w_2	$EFCADBG$
w_3	ECF
w_4	EFA
w_5	E

TABLE 4.2: Sport dataset: consensus rankings

The consensus estimated for each weighting vector is shown in Table 4.2. With the first two weighting vectors, we obtain the same consensus, which corresponds to the one obtained by Amodio *et al.* (2016). With the other 3 weighting structures we obtained different consensuses, and with w_5 2 different solutions corresponding to the same value of τ_x^w, always with the course "E" in first position.

The second real data example is reported by Emond and Mason (2000). In this dataset, 112 experts were asked to rank 15 Departmental initiatives

(A to Q), competing for limited funds. In this example, some ties occur, and some rankings are not complete. This dataset has also been analyzed by Amodio et al (2016). Table 4.3 shows the results corresponding to the five weighting structures (conceptually equivalent to the ones shown in the Table 4.1, but extended to 15 items). The first consensus, as expected, is the one found by Emond and Mason, while it changes as soon as the weighting structure changes, even if D is always in first position.

w_1	$DL < EM >< ABP >< CN > IHFG < OQ >$
w_2	$DEMF < ABL > IP < CN > HG < OQ >$
w_3	$DEMPF < AL >$
	$D < EM > P < AL >$
w_4	$DEM < AFL >$
w_5	D

TABLE 4.3: Emond and Mason dataset: consensus rankings

4.7 Concluding remarks

We provided a weighted rank correlation coefficient τ_x^w for linear and weak orderings, as an extension of τ_x^w for linear orderings. We demonstrated the correspondence between τ_x^w and the weighted Kemeny distance and, finally, we showed that, in the case of tied rankings and $w_i = \frac{1}{m-1}$ for all i, the weighted rank correlation coefficient τ_x^w is equal to the Emond and Mason rank correlation coefficient τ_x . From the simulation study, we demonstrated that the modified BB algorithm allows us to find the true consensus and to verify the effect of the weighting vector. The analysis of the two real datasets shows, as demonstrated analytically, that with $w_i = \frac{1}{m-1}$ for all i we obtain the same solution without weightings, while the solutions always differ as soon as we simplify the weighting structure.

Some crucial points could represent the basis for future developments: first of all, to take into account the multiple solutions of the consensus process, since only one random solution has been considered for now (in order to facilitate the implementation process); then, the optimization of the implemented procedures in order to achieve faster algorithms; in the end, the development of the same analysis for item positions for a complete consensus process.

Chapter 5

Decision trees for positions'weighted ranking data

5.1 Introduction

Decision trees are a competitive tool with respect to other state-of-the-art methods in terms of predictive accuracy and, not less relevant, they are generally considered as being more comprehensible and interpretable than most other model classes (Cheng *et al.*, 2009). Piccarreta (2010) proposes binary trees for dissimilarity data, listing preference data too as a possible practical application, but the paper does not consider dissimilarities that are specific to preference data. The class of distance-based ranking models states that the probability of observing a specific ranking depends on the distance between the observed ranking and the modal ranking, i.e.: "they assume a modal ranking and the probability of observing a ranking is inversely proportional to its distance from the modal ranking. The closer to the modal ranking, the more frequent the ranking is observed" (Lee and Yu, 2010). A further development in the distance-based approach is due to Lee and Yu (2010) who developed a distance-based tree model introducing a weighted distance, where the weights are linked to items in the preference ranking.

The possibility of using weights can take into account crucial concepts, totally ignored by classical non-weighted distances: item relevance and positional information. As a matter of fact, in some real situations, it might be relevant to give more importance to changes in the top-positions of rankings or it might be more important to emphasize changes occurring in highly-relevant items rather than changes in less important items. Distance measures with either position or element weights are well discussed by Kumar and Vassilvitskii (2010).

The aim of this chapter is to introduce weighted distance-based tree models, inspired by Plaia and Sciandra (2019).

Decision trees represent a simple non-parametric statistical methodology designed, originally, for classification and prediction issues. Their main feature consists that the space spanned by all predictor variables is recursively partitioned into a set of rectangular areas, such that observations with similar response values are grouped and, at the end of the recursive procedure, the resulting groups are as homogeneous as possible with respect to the considered response.

The Classification and Regression Tree (a.k.a. CART) procedure has been described for the first time by Breiman *et al.* (1984) and it is, definitely, the most popular decision tree methodology and widely used in multidisciplinary fields. Nevertheless, a variety of algorithms have been proposed in the literature for the construction of decision trees, for categorical, discrete or continuous response, such as C4.5, CHAID or QUEST (see Philip *et al.* (2010) for some references). CART is mainly characterized by two stages: a growing phase and a pruning one.

The CART procedure, starting from the root node containing the whole dataset (or only a learning sample), uses a recursive partitioning algorithm in order to create a tree where each node t represents a subset of the partition. In the binary tree structure, all internal nodes have two child nodes, whereas the nodes with no descendants are called leaf nodes or terminal nodes. At each step of the partitioning process, a splitting rule is used to split the $N(t)$ objects in node t into a left and a right node. The splitting process is an important step, the core issue being how to choose the *splitting rule*, the rule that performs the splitting of the sample into smaller parts. Generally, the splitting rule is chosen on the basis of the quality-of-split criterion, which is equivalent to choosing a split among all possible splits at each node so that the resulting child nodes are the "purest", where pureness is meant in terms of homogeneity (with respect to the response) of observations in the same node. Maximum homogeneity of child nodes is defined by a so-called *impurity function* $i(t)$, a generic function satisfying the following three properties: (a) it is minimum when the node is pure; (b) it is maximum when the node is the most impure; (c) its value does not change if items are renamed. Specifically, a splitting criterion is based on the reduction in impurity resulting from the split s of node t, with the best split chosen as the one maximizing the impurity reduction (Shih, 2001). In the end, once the tree has been built, terminal nodes must be associated with a predicted response: for classification trees (categorical response variable) the assignment is based on a simple majority rule, while for a regression tree (quantitative response),

a simple mean of the response variable of the objects in the node can be considered. The problem of "class assignment" for preference data will be discussed in Section 5.2.1.

5.2 Decision trees for preference data

The weighted Kemeny distance measure, introduced by García-Lapresta and Pérez-Román (2010) and reported in Equation (2.10), represents the main ingredient for the definition of our proposal, which consists of extending classical splitting criteria, used in decision tree modelling, in order to include other measures that can be used to evaluate impurity when faced with preference data.

5.2.1 Splitting criterion and impurity function for preference data

The main problem in extending univariate regression trees to the multivariate response case deals with generalizing the definition of the partitioning metric. In order to avoid the problems linked to the multivariate nature of the ranking data, in this work the vectors of preferences will be treated as a unique multivariate entity. Specifically, when the m items are completely ranked, the ranking process can be seen as a permutation function from $\{1, \ldots, m\}$ onto $\{1, \ldots, m\}$. Hence, if a label is assigned to each permutation defined from a set of distinct items, each ranking vector can be identified through a label that, used as a response variable, allows applying the classical univariate classification tree methodologies. The recursive binary partitioning process used by the CART methodology, starting from the root node, produces a nested sequence of subtrees

$$T_k = \{root - node\} \subset \cdots \subset T_0 = \{full - tree\}$$

obtained, at each step, through a splitting criterion which works with the maximization of $\Delta i(s,t)$, i.e.: the reduction in the impurity resulting from the split s in node t. In a line:

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

where p_L and p_R are the proportions of units in node t falling in the left child node t_L and in the right child node t_R, respectively, at the s-th split.

When the impurity has to be evaluated working with ranking data, the impurity function can be properly modified as follows:

$$i(t) = \sum d(a, b), \tag{5.1}$$

where d is the chosen distance measure between the orderings a and b and summation is extended to all the couples of rankings in the node t. A decrease in node impurity at each step will be evaluated according to all covariates and respective split points. Piccarreta (2010) proposes an impurity function based on dissimilarities, without proposing specific distances for preference data. In order for the impurity function to take into account the ordinal nature of rankings, we propose to use the weighted Kemeny distance (Equation 2.10) , as distance d in Equation (5.1),

According to the CART methodology, tree optimization implies cutting off insignificant nodes through a pruning process based on some cost-complexity criteria that will lead obtaining the optimum tree size. This pruning procedure is out of the aim of this work, but classical cost-complexity pruning procedure can be applied.

5.2.2 Rank aggregation in the leaves

Once the tree has grown, the definition of a consensus ranking is necessary for assigning a class label or class ranking to each node, i.e. the ranking in the best agreement with all the rankings in the node (D'ambrosio *et al.*, 2015a). Among the several consensus ranking measures proposed in the literature, the median ranking approach will be used in the examples presented in the next section. The median is defined as the ranking corresponding to the minimum sum of the distances of all rankings from it or it is, equivalently, the ranking that maximizes the average τ_x^w rank correlation coefficient between itself and the other rankings belonging to the set of rankings (D'Ambrosio and Heiser, 2016). To compute the median ranking we use the adapted versions of the Branch and Bound algorithm, proposed in Sections 3.5 and 4.5 for positional weighted ranking data, without and with ties, respectively.

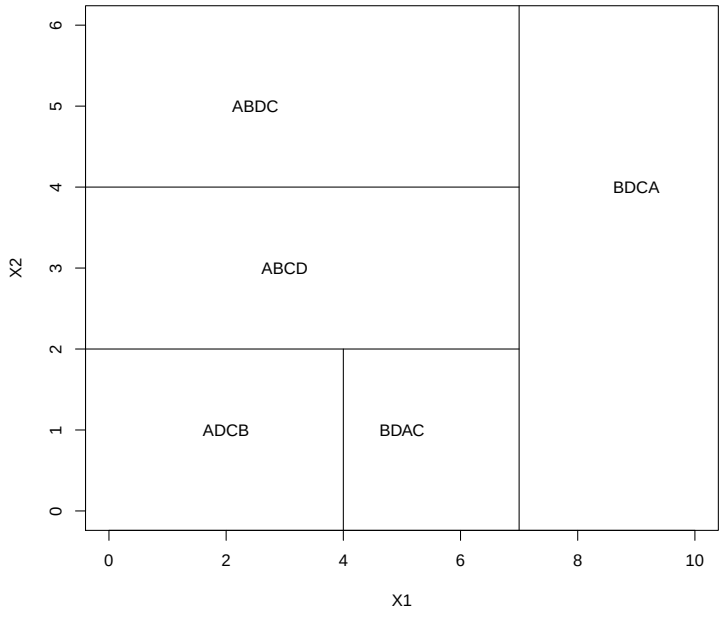
The ranking having the highest correlation with the entire set of the n rankings is the consensus ranking, i.e. the most representative ordering among all possible solutions. Hence, given a $n \times m$ matrix $\mathbf{X}$, where the l-th row represents the ranking associated to the l-th judge, the consensus ranking, i.e. the ranking r (a candidate within the universe of the permutations with/without replacement of m elements) that best represents

the matrix $\mathbf{X}$, is that ranking that maximizes $\sum_{l=1}^{n} \tau_x^w({}_l x, r)$, see Equations (3.5) and (4.4).

The implementation of the decision trees above described has been possible through the adaption of a proper splitting function in the "rpart" package (Therneau *et al.*, 2010).

5.2.3 Simulation study

We are interested in evaluating the effect both on the splits and on the leaf labels of different weighting vectors w. For this reason, following D'Ambrosio and Heiser (2016), we consider a theoretical population partition of the predictor space (X_1 and X_2) reported in Figure 5.1, with $X_1 \sim U(0, 10)$ and $X_2 \sim U(0, 6)$.



.

FIGURE 5.1: Theoretical partition of the predictor space: $X1 \sim U(0, 10)$, $X2 \sim U(0, 6)$

The number of rankings falling in each group was defined by a random number drawn from a normal distribution $N \sim (10, 2)$ and each number was divided by the summation of all of them, obtaining a relative frequency distribution for each sub-partition. The rankings of 4 items

of each sub-partition were generated from a Mallows Model (see Equation (3.9)), varying the dispersion parameter θ. The central permutations used for generating the rankings related to each sub-groups are shown in Figure 5.1.
Considering three levels for the sample size (100 and 300) and three different level of noise (low with $\theta = 50$, medium with $\theta = 2$ and high with $\theta = 1$), the experimental design counts $2 \times 3 = 6$ different scenarios. Figure 5.2 shows one of the nine datasets considered in the simulation study, obtained with $\theta = 50$ and $n = 300$.

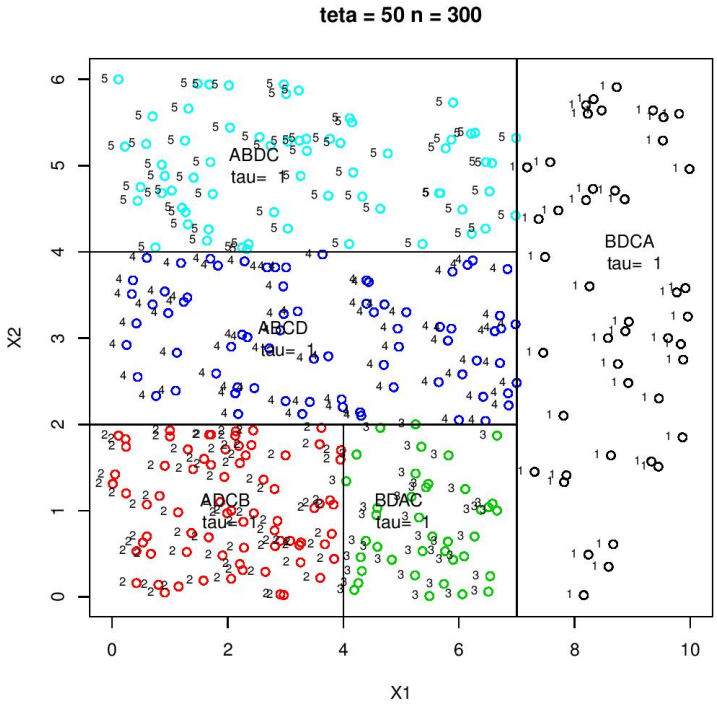
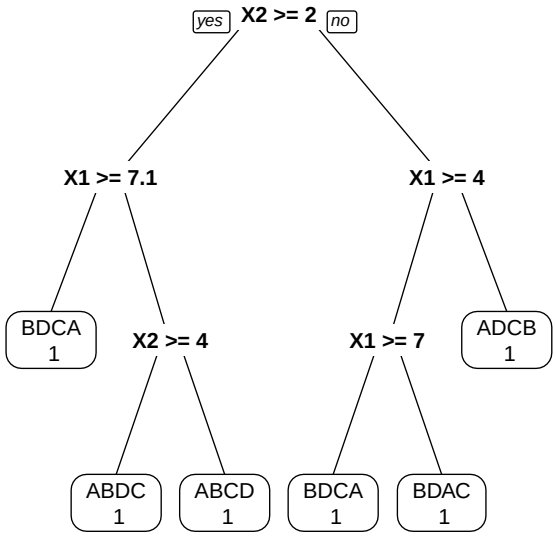


FIGURE 5.2: Generation of homogeneous groups of ranking from the theoretical partition with: $X1 \sim U(0, 10)$ and $X2 \sim U(0, 6)$, $theta = 50$ and $n = 300$

For each dataset five different weight vectors are considered : $w_1 = (1/3, 1/3, 1/3)$, $w_2 = (3/6, 2/6, 1/6)$, $w_3 = (1/2, 1/2, 0)$, $w_4 = (2/3, 1/3, 0)$ and $w_5 = (1, 0, 0)$.
With reference to the data in Figure 5.2, Figure 5.3 reports two of the five trees obtained: in particular, (A) shows the tree corresponding to w_1, which perfectly recreates the original partition of the predictor space;

(B) corresponds to w_3 and, as expected, does not perform the two splits $X \gtreqless 4$ and $X \gtreqless 7$ (the couples of rankings below each of the split in (A) do not differ for the first two positions). In fact, $w_3 = (1/2, 1/2, 0)$ means that we are not interested in positions 3 and 4, i.e. the τ_x^w between two rankings that differ only for the positions 3 and 4 is maximum ($= 1$).



(A)

For each dataset generated with $n = 300$ and $\theta = (1, 2, 50)$, τ_x^w was measured before and after applying the decision trees with all five weights' vectors and the distributions obtained are shown in Figures 5.4, 5.5 and 5.6. The higher the homogeneity among rankings, the better the ability of the decision trees to detect groups of similar rankings according to different values of $X1$ and $X2$. When rankings have an initial high homogeneity ($\theta = 50$, see Figure 5.6) the trees show an almost perfect performance for all kinds of position weights, except for the case of w_5: assigning a weight only to the first position penalizes the performance of the trees. This penalization reduces as the homogeneity tends to decrease ($\theta = 1$, see Figure 5.4).

5.3 Conclusion

In this chapter, we have focused on distance-based decision trees for ranking data, when the position occupied by items is relevant. We have proposed the weighted Kemeny distance as impurity function and the

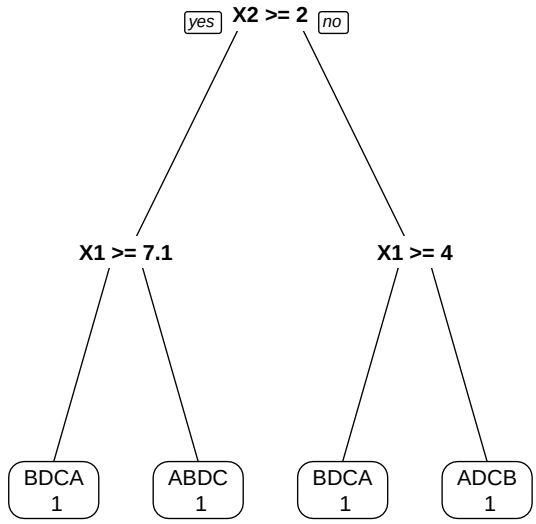


FIGURE 5.3: (B)
Decision tree models for weighted rankings with weights vector w_1 in (A) and w_3 in (B).

relative proper weighted correlation coefficient in order to achieve the consensus measure in the terminal nodes. Our methodology found to be capable of identifying correctly homogeneous groups of rankings when more than one position is taken into account. The implementation of a faster algorithm for the rpart package (Therneau *et al.*, 2010) could lead to work faster in the presence of a large number of items. Further developments could be, hence, a replication of the same analyses with an increasing number of items.

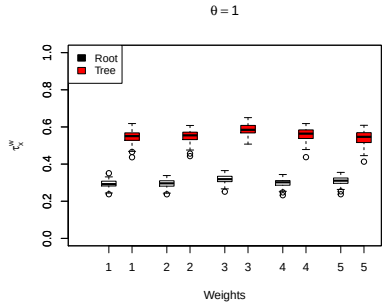


FIGURE 5.4: Measure of τ_x^w both in the root and overall tree ($theta = 1$ and $n = 300$)

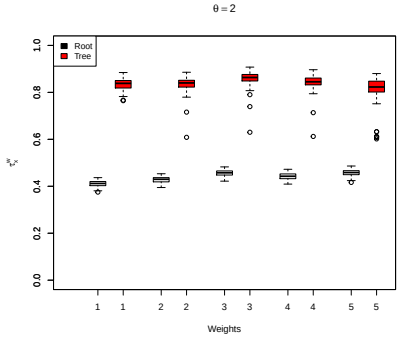


FIGURE 5.5: Measure of τ_x^w both in the root and overall tree ($theta = 2$ and $n = 300$)

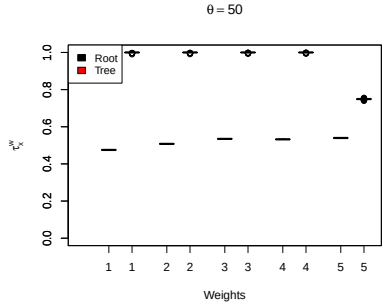


FIGURE 5.6: Measure of τ_x^w both in the root and overall tree ($theta = 50$ and $n = 300$)

Chapter 6

Ensemble methods for position weighted ranking data: a new proposal

6.1 Introduction

In order to detect which predictors better explain a phenomenon, a huge use of decision trees has been observed in the past. These techniques are quite intuitive but they are unstable: small perturbations bring big predictive changes. An approach used to make decision more reliable is to combine the output of multiple trees, leading to a more stable procedure called ensemble method. This technique has been applied mostly to numeric prediction problems and classification tasks. In the last years, some attempts to extend the ensemble methods to ordinal data can be found in the literature, but not a concrete methodology has been provided for preference data. The interest of this research lies in building decision trees and, consequently, ensemble methods for ranking data. Boosting and Bagging are two of the most known ensemble methods and this work proposes a theoretical and computational definition of them for ranking data.

In the 1980s Breiman *et al.* (1984) developed Classification and Regression Trees (CART) as alternative non-parametric approaches to classification and regression. The resulting tree-structured predictors are simply obtained as functions of the input variables, but they are unstable, i.e. small perturbations in the training set could bring to large changes in the predictive results. Unstable classifiers such as decision trees have high variance and low bias. Breiman (1996) suggested to improve the accuracy of decision trees using the bootstrapping technique, a general statistical method in which several (nondisjoint) training sets are obtained by

drawing randomly, with replacement, from a single base dataset. Therefore perturbing the training set and then combining the multiple decision trees into a single predictor, it is possible to improve the prediction process. It is a quite simple idea: many bootstrap samples are drawn from the available data, some prediction process is applied to each bootstrap sample, and then the results are somehow combined. Breiman labelled this methodology P&C (perturb and combine) . One of the most known P&C methods is Bagging (Boostrap AGGregatING), which perturbs the training set several times in order to generate multiple predictors and combine them by simple voting and by averaging in classification and regression, respectively. The bootstrap samples drawn at each iteration have the same probability to be drawn and that guarantees independent classifiers.

Freund *et al.* (1996) and Freund and Schapire (1998), worked on the P&C algorithms to increase the rate of the training set error going to zero: they increased the probability to be drawn in next iterations for those units mostly misclassified in the previous steps. They introduced therefore another perturb and combine method called Boosting. Unlike bagging, it's clear that the samples used in the several iterations are not independent. The Boosting procedure is considered by Breiman *et al.* (1998) as a special case of the class of *arcing* (adaptive resampling and combining) classifiers, where an increased probability to be drawn in next iterations is assigned to units frequently classified as incorrect. Bagging and Boosting have been widely applied to quantitative or qualitative response but, up to our knowledge, no extension to cope with preference data exists in literature.

The purpose of this chapter is to define, analytically and empirically, the Boosting and Bagging algorithms for rankings, with or without ties. The algorithms will be applied to real and simulated data.

6.2 Emsemble methods for ranking data without weights

"The idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models" (Hastie *et al.* (2005)). In this part of the Thesis, Ensemble methods are built using trees as building blocks in order to construct more powerful prediction models, than a single tree. Ensemble learning can be broken down into two tasks: developing a population of base learners (decision trees) from the training data, and then combining them to form the composite predictor.

In this section the first point is going to be discussed and in Section 6.2.4 the way of aggregating the trees will be shown.

6.2.1 Bagging algorithm with replacement

The simplest implementation of the idea of generating quasi-replicate training sets Breiman (1996) is:

- 1. Start with $w_b(i) = 1/n; \forall i = 1, 2, \ldots, n$
- 2 Repeat for $b = 1, 2, \ldots, B$
  - a Fit the classifier $C_b(x_i)$

In a few words, several training datasets of the same size are chosen at random, with replacement, from the training set. The decision tree is built for each of them, leading to a proper prediction of ranking for each unit. Once the trees are built, a final ranking prediction is assigned to each unit through the consensus ranking process described in Sections (3.5) and (4.5). Combining multiple classifiers decreases the expected error by reducing the variance component. The more the classifiers included, the greater the reduction in error.

6.2.2 AdaBoost.M1 algorithm for rankings

Boosting is a method that combines classifiers, which are iteratively created from weighted versions of the learning sample, with the weights adaptively adjusted at each iteration to give a higher probability to be drawn to misclassified cases in the previous step. The final predictions are obtained by weighting the results of the iteratively produced predictors. There are many versions of boosting algorithms, but the most used is the AdaBoostM1 (Freund *et al.* (1996)) and it is specifically used for classifications.

In a few words, the boosting technique consists of repeatedly using the base weak learning algorithm, on differently weighted versions of the training data, leading to a sequence of weak classifiers that are finally combined someway. A weak learner is a learning algorithm which provides classifiers with a probability of error less than that of simply random guessing (0.5, in the binary case). A strong learner, instead, is capable (given enough training data) to lead to classifier algorithms with small error probability.

Firstly, the AdaBoost.M1 algorithm adapted to position weighted rankings is described as follows. The weight $w_b(i)$ is assigned to each i-th observation and it's initially equal to $1/n$ for all instances. This value is updated after each step. A basic classifier, $C_b(xi)$, is built on the new training set (T_b) and is applied to every training example. For fitting C_b the weights w_b are used for drawing a random sample S_b of the data, and then C_b is used for learning from S_b. The error of this classifier is represented by e_b and is calculated differently from the classification problems, where the fact that a label is classified wrongly or correctly is the relevant point. In the field of ranking data, what counts mainly is given by how far the predicted ranking is away from the real one. That's why the error measure used in each classifier is based on the rank correlation coefficient:

$$e_b = \sum_{i=1}^{n} w_b(i) \left[1 - \frac{\tau_x(i) + 1}{2}\right]. \tag{6.1}$$

From the error of the classifier in the b^{th} tree, α_b is computed for updating the weights related to each tree. Recalling Freund and Schapire: $\alpha_b = \ln((1 - e_b)/e_b)$. Anyway, the new weight for the $(b + 1)$-th tree is $w_{b+1}(i) = w_b(i) \exp \alpha_b \tau_x(i)$ and it's normalized after the computation on all the trees.

The bigger the distance between the ranking associated to an observation and the original ranking, the higher the probability that this observation is resampled in the new iteration. The sequence of trees tries to correctly identify the rankings, focusing more on those hardly predictable in the right way. When the classifiers'error decreases more and more then a high accuracy in prediction is reached. The iterative procedure continues until the stopping criterion (i.e. $\alpha_b > 0.5$ or the maximum number of trees) is reached.

- 1. Start with $w_b(i) = 1/n; \forall i = 1, 2, \ldots, n$
- 2 Repeat for $b = 1, 2, \ldots, B$
  - a Fit the classifier $C_b(x_i)$ using weights $w_b(i)$ on T_b
  - b Compute: $e_b = \sum_{i=1}^{n} w_b(i)\left[1 - \frac{\tau_x(i)+1}{2}\right]$ where $\tau_x(i) = \tau_x(\hat{y}_i, y_i))$ and $\alpha_b = \frac{1}{2} \ln((1 - e_b)/e_b)$
  - c Update the weights $w_{b+1}(i) = w_b(i) \exp \alpha_b \tau_x(i)$ and normalize them

---

6.2.3 Bagging algorithm with OOB

As an alternative to Bagging with replacement, we can consider for the Bagging method the out-of-bag (OOB) approach. This approach consists in fitting the trees on around two-thirds of the observations in the training set. The ability of the ensemble method to predict as better as possible the response variable is measured using the one-third of the observations left out in the fitting process. In order to obtain a single prediction for a generic unit belonging to the OOB set, usually, an average value is taken among all the numeric predictions provided by all the fitted trees if the regression is the case, otherwise, in the classification case, the simple majority vote is followed. Once all the OOB units are predicted an overall OOB mean squared error (in regression problems) or a classification error (in classification issues) is computed. In the particular case we deal with, it's necessary to take into account the nature of the response variable: ranking data. Hence, the OOB error resulting is obtained working with an average rank correlation coefficient between the real rankings and those predicted.

In order to reach a higher reduction in error, we applied Bagging sampling the bootstrapped trainings without replacement. Even if it's true that replacement leads to higher variability (with respect to the original set of data), it's also true that its effect is less as the size of the training set increases. Since in our application we consider large dataset, we consider the possibility to sample the units belonging to each training set, tree by tree, without replacement.

6.2.4 Rank aggregation and test error measurement

In order to assign a predicted ranking to each unit, the rank aggregation process explained in Sections 3.5 and 4.5 is used, after building all the trees in the ensemble procedure. The partial ranking for a generic i-th observation, $\hat{y}_{ib}$ (see Table 6.1), will be obtained as follows:

$$\hat{y}_{ib} = \arg\max \sum_{k=1}^{b} \alpha_k \tau_{x,k}(i), \qquad b = 1, 2, \dots B \tag{6.2}$$

where α_k, the weight related to the k-th tree, is equal to $\frac{1}{2}\ln((1-e_k)/e_k)$ (Breiman *et al.* (1998)) in the Boosting methodology and equal to 1 in the case of Bagging. Furthermore, $\hat{y}_{iB}$ represents the final ranking for the i-th unit.

Weights	a_1	a_2	...	a_b	...	...	a_B
Obs	**tree**$_1$	**tree**$_2$	...	**tree**$_b$	...	...	**tree**$_B$
1	$\hat{y}_{11}$	$\hat{y}_{12}$	...	$\hat{y}_{1b}$	...	...	$\hat{y}_{1B}$
2	$\hat{y}_{21}$	$\hat{y}_{22}$	...	$\hat{y}_{2b}$	...	...	$\hat{y}_{2B}$
.	.	.	...	.	...	...	.
.	.	.	...	.	...	...	.
.	.	.	...	.	...	...	.
n	$\hat{y}_{n1}$	$\hat{y}_{n2}$	...	$\hat{y}_{nb}$	...	...	$\hat{y}_{nB}$
Error	err(1)	err(2)	...	err(b)	...	...	err(B)

TABLE 6.1: Predicted rankings per tree

Once each unit has been assigned a final ranking tree by tree (See Equation (6.2)), the error assigned to each tree is computed as:

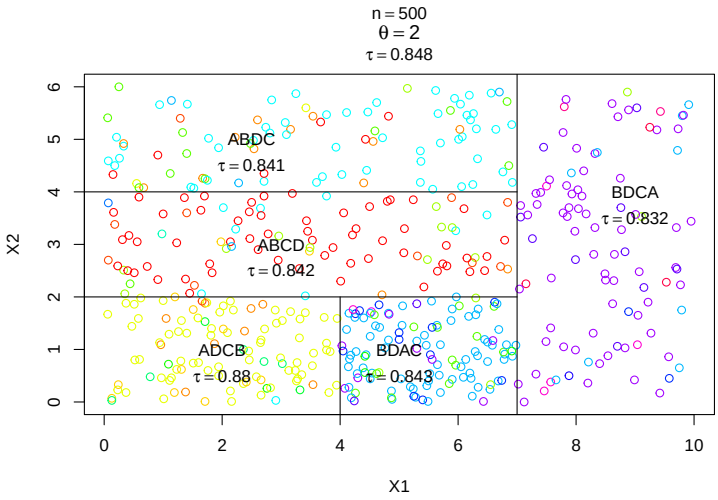
$$err(b) = 1 - \frac{\tau_x(b) + 1}{2}, \tag{6.3}$$

where $\tau_x(b)$ is the average of τ_x of the b-th tree over all the units in the b-th tree.

6.2.5 Real example and a simulation experiment

In order to evaluate the performance of the Ensemble methods described above, we performed a simulation study and the application on a real dataset. To this aim, following D'Ambrosio and Heiser (2016), we considered a theoretical population partition of the predictor space (X_1 and X_2), with $X_1 \sim U(0, 10)$ and $X_2 \sim U(0, 6)$. The number of rankings of 4 items falling in each group was defined by a random number drawn from a normal distribution $N \sim (10, 2)$ and each number was divided by the summation of all of them, obtaining a relative frequency distribution for each sub-partition. The rankings of each sub-partition were generated from a Mallows Model using the PerMallows R package (Irurozki *et al.*, 2016), described in Section 3.6.

In our simulation, we generated rankings assuming the Kemeny distance and varying the dispersion parameter θ, according to three different level of noise (low with $\theta = 2$, medium with $\theta = 0.7$ and high with $\theta = 0.4$). Considering two levels for the sample size (200, 500) the experimental design, hence, counts $3 \times 2 = 6$ different scenarios (three levels



.

FIGURE 6.1: Empirical partition of the predictor space, generating high homogeneous groups of rankings ($\theta = 2$), with $n = 500$.

of noise and two sample sizes). Figure 6.1 shows one of the six datasets considered in our experiment.

We applied the ensemble methods defined in Section 6.2 to all the six scenarios, fixing the number of trees to 300. Looking at the errors produced by Boosting (Figure 6.2), the methodology is able to perform very well when there is a high level of heterogeneity among the rankings ($\theta = 0.4$)

The training error and the test error for $n = 500$ are almost costant after 100 trees. Worst performances are visible for $n = 200$ compared to the case with $n = 500$. Moreover, the higher the heterogenity among the rankings, the higher the ability of the boosting method.
Bagging performs worse than Boosting. In particular, as concerns the Bagging approach with OOB (Figure 6.4), it's pretty clear that the method is unstable. The training error was measured with the OOB approach and the random absence of one third of the predicted rankings, probably, leads to jagged errors with a zig-zag shape around the initial error. On the same simulated data we applied Bagging (with replacement) and Bagging (with OOB and without replacement) and the outcomes obtained are reported in Figure 6.3 and in Figure 6.4, respectively. As concerns the first case, it was enough stopping the procedure at 100 trees, since the errors showed to be quite stable. In all the scenarios, Bagging errors

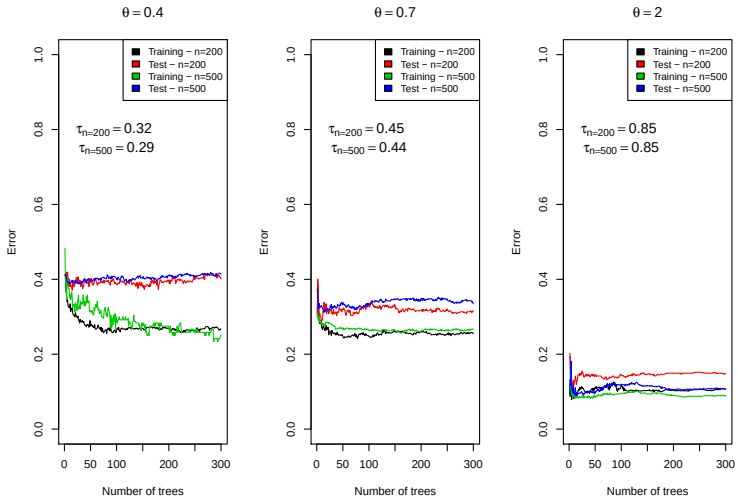
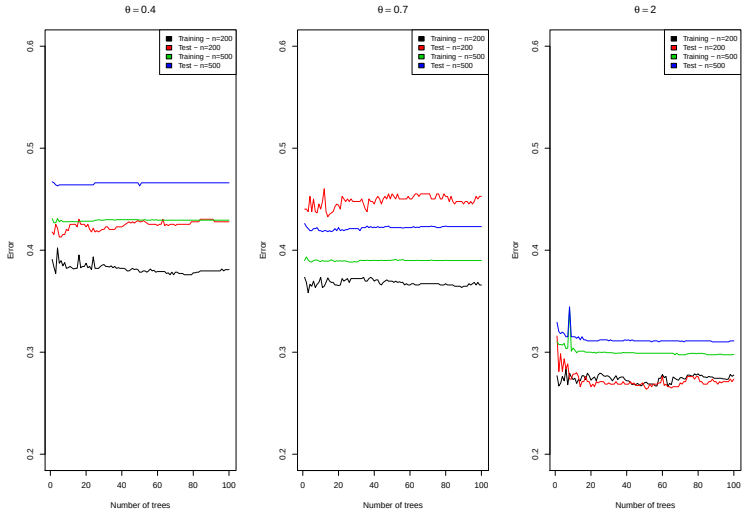


FIGURE 6.2: Boosting for all the simulated scenarios: different levels of homogeneity among the rankings, $\theta = (0.4, 0.7, 2)$, and two sample sizes, $n = (200, 500)$

decreases when the heterogeneity decreases, independently on the data size. Bagging with OOB is insensitive to the data size too, but the procedure results instable. In order to analyze if the procedure is sensitive to the number of splits in each tree, i.e. to its depth, we considered for the Boosting method (see Figure 6.5), a maximum number of splits first equal to 2 and then 4. It becomes clearer, going from $\theta = 0.4$ to $\theta = 2$, that the mean error computed using 4 splits (both in the training and in test sets) tends to be lower than that with the number of split equal to 2.

Real case application

The two ensemble methods have been applied to a real case (Figures 6.6 and 6.7). The real application is related to the dataset "vehicle" available in the UCI repository (*https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/vehicle/*). This dataset concerns 846 units, 18 variables and four items and we fixed the number of trees to 350. Figure 6.6 compares the outcomes obtained applying Boosting and Bagging (with replacement) to the data: the boosting shows a better performance again. The two kinds of error become stable after 100 trees. Figure 6.7 compares, instead, the outcomes obtained applying Boosting and Bagging (with OOB and without replacement). Bagging procedure show results similar to the simulation's ones: the training error doesn't apport a good



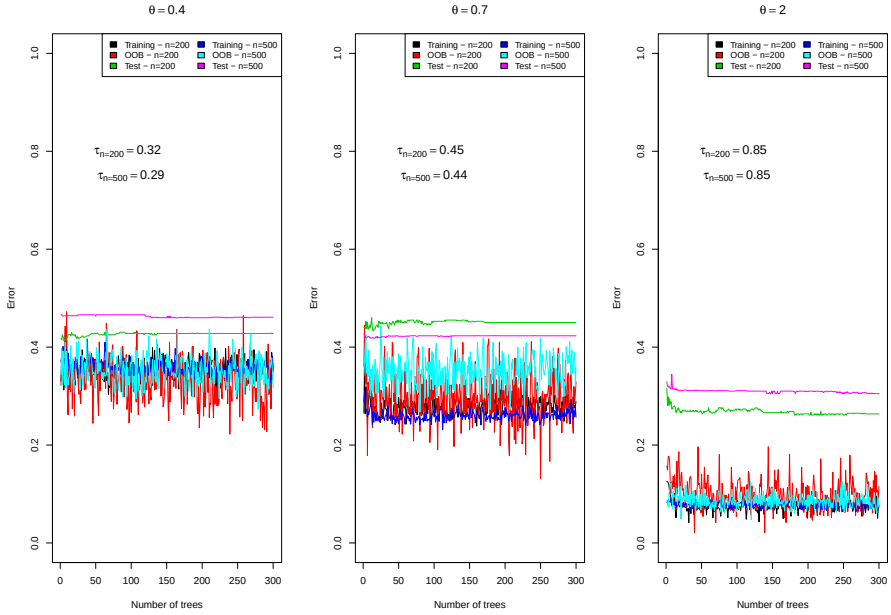
.

FIGURE 6.3: Bagging for all the simulated scenarios with 100 trees: different levels of homogeneity among the rankings, $\theta = (0.4, 0.7, 2)$, and two sample sizes, $n = (200, 500)$

performance, since its values are quite unstable and move around 0.06. The same conclusion concerns the OOB error, which is even more unstable. The test error after an initial decreasing becomes stable around 0.09 after 50 trees.

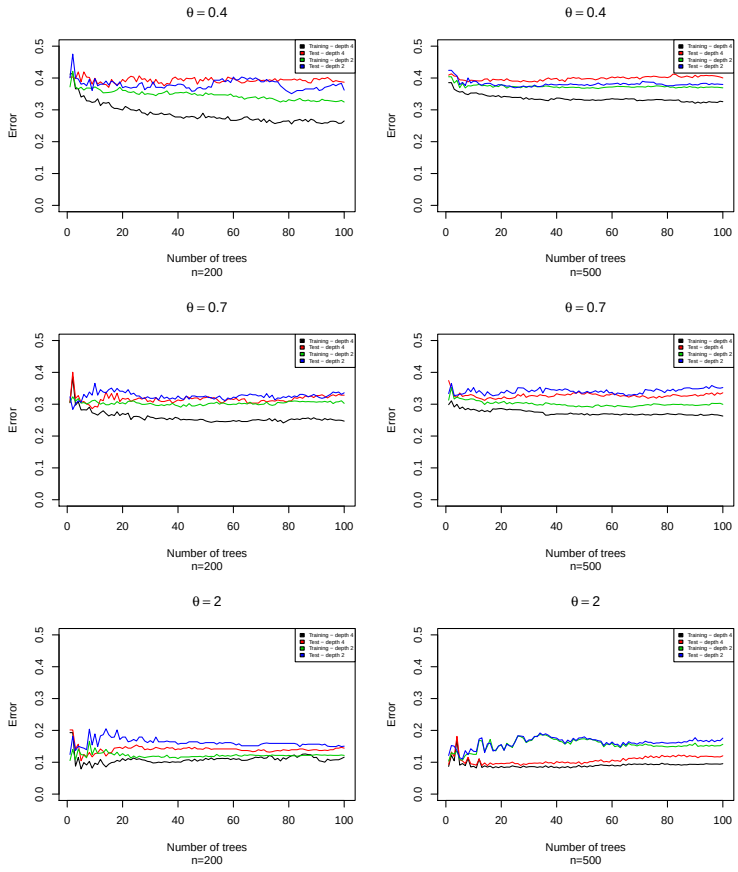
6.2.6 Conclusion

In this section we propose two ensemble methodologies, Boosting and Bagging (with replacement and with OOB, without replacement), for ranking data. Looking at the single decision tree, we used the Kemeny distance (Kemeny Snell) as a measure of impurity in the splitting process and its related rank correlation coefficient (τ_x proposed by Emond and Mason) for identifying the median ranking in the final nodes. Once the trees are built up, τ_x was employed for assigning the median ranking as the final prediction, tree by tree, and for measuring the relative error. We applied the above methodologies to simulated data and to a real case showing that boosting outperforms bagging (both with and without replacement). By means of simulations the sensitivity of the procedures to the number of threes, the heterogeneity of the data and the depth of the single tree has been studied.



.

FIGURE 6.4: Bagging for all the simulated scenarios: different levels of homogeneity among the rankings, $\theta = (0.4, 0.7, 2)$, and two sample sizes, $n = (200, 500)$

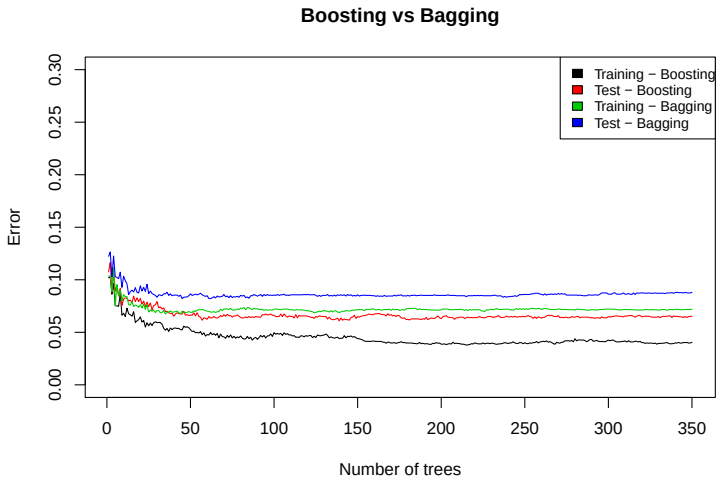


.

FIGURE 6.5: Boosting built up for each simulated dataset, using two depths for the trees (2 and 4)

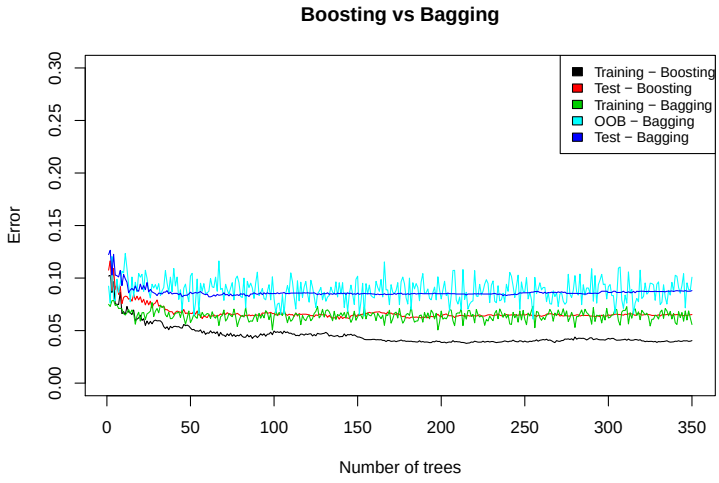
6.3 Emsemble methods for ranking data with positional weights

In this part of the Thesis, the well known Ensemble method called AdaboostM1 is built using trees as building blocks in order to construct more powerful prediction models, rather than a single tree, when the position occupied by the items is important. Hence, the two tasks relevant for an Ensemble methodology (i.e. developing a population of base learners from the training data, and then combining them to form the composite predictor) are hereby adapted to the case of position weights assigned to the items. Bagging outcomes are too unstable in the case of rankings



.

FIGURE 6.6: Boosting and Bagging applied to Vehicle dataset



.

FIGURE 6.7: Boosting and Bagging applied to Vehicle dataset

without position weights (see Section 6.2.5) which is a subcase of position weights (uniform weights, see Section 4.4). For this reason, the focus in this section is on adapting only Boosting in the case of rankings with position weights.

6.3.1 AdaBoost.M1 algorithm for rankings with weights

The AdaBoost.M1 algorithm has the same structure introduced in Section 6.2.2, but since the focus is on the items'positions the rank correlation coefficient τ_x introduced by Emond and Mason (2002) is suitably replaced by its weighted version τ_x^w (Plaia *et al.*, 2019b) and (Plaia *et al.*, 2018a) (See Equations (3.1) and (4.1)). The error measure (Equation (6.1)) used in each classifier is based now on the weighted rank correlation coefficient which takes into account the positions of items:

$$e_b = \sum_{i=1}^{n} w_b(i) \left[1 - \frac{\tau_x^w(i) + 1}{2} \right]. \tag{6.4}$$

The main concepts is the same: the bigger the distance between the ranking associated to an observation and the original ranking, the higher the probability that this observation is resampled in the new iteration. Hence, the sequence of trees tries to correctly identify the rankings, focusing more on those hardly predictable in the right way. The iterative procedure continues until the stopping criterion (i.e. $\alpha_b > 0.5$ or the maximum number of trees) is reached. The pseudo-code of Section 6.2.2 is proposed, replacing τ_x with τ_x^w:

- 1. Start with $w_b(i) = 1/n; \forall i = 1, 2, \ldots, n$
- 2 Repeat for $b = 1, 2, \ldots, B$
  - a Fit the classifier $C_b(x_i)$ using weights $w_b(i)$ on T_b
  - b Compute: $e_b = \sum_{i=1}^{n} w_b(i) \left[1 - \frac{\tau_x^w(i)+1}{2}\right]$ where $\tau_x^w(i) = \tau_x^w(\hat{y}_i, y_i))$ and $\alpha_b = \frac{1}{2} \ln((1 - e_b)/e_b)$
  - c Update the weights $w_{b+1}(i) = w_b(i) \exp \alpha_b \tau_x^w(i)$ and normalize them

---

6.3.2 Rank aggregation and test error measurement

In order to assign a predicted ranking to each unit, the rank aggregation process explained in Section 6.2.4 is used, after building all the trees in the ensemble procedure. The final ranking for a generic i-th observation, $\hat{y}_{iB}$, will be obtained as follows:

$$\hat{y}_{iB} = \arg\max \sum_{b=1}^{B} \alpha_b \tau_{x,b}^{w}(i), \tag{6.5}$$

where α_b, the weight related to the b-th tree, is equal to $\frac{1}{2}\ln((1-e_b)/e_b)$ (Breiman *et al.* (1998)) in the Boosting methodology and 1 in the case of Bagging.

Following the same procedure described in Section 6.2.4, once each unit has been assigned a final ranking tree by tree (See Equation (6.5)), the error assigned to each tree is computed as:

$$err(b) = 1 - \frac{\tau_x^w(b) + 1}{2}, \tag{6.6}$$

where $\tau_x^w(b)$ is the average of τ_x^w of the b-th tree over all the units in the b-th tree.

6.3.3 Real example and a simulation experiment

For the simulation, we repeated what described in Section 6.2.5. Moreover, we introduced 5 different structure of weights: $w_1 = (1/3, 1/3, 1/3)$, $w_2 = (3/6, 2/6, 1/6)$, $w_3 = (1/2, 1/2, 0)$, $w_4 = (2/3, 1/3, 0)$ and $w_5 = (1, 0, 0)$. In other words, equal and decreasing weights were considered, at first involving all the weights and then only half of the total amount of the vectors, and, finally, only the first position was weighted. The sample size used for all the datasets was 300. The experimental design, hence, counts $3 \times 5 = 15$ different scenarios (three levels of noise and five different weighted vectors).

We applied the Boosting method defined in Section 6.3.1 to all the scenarios, fixing the number of trees to 300. Looking at the errors produced (Figure 6.8), the methodology is able to perform very well when there is a high level of heterogeneity among the rankings ($\theta = 0.4$)

It's clear that the method is stable and the items'position weights have a greater effect on the error as the heterogeneity among rankings increases ($\theta = 0.4$). When only the first position is considered ($w_5 = (1, 0, 0)$) the error outcome is unstable with a high level heterogeneity

($\theta = 0.4$) and it becomes stable as θ increases but once again we find out that this kind of weights penalizes the performance of the trees.

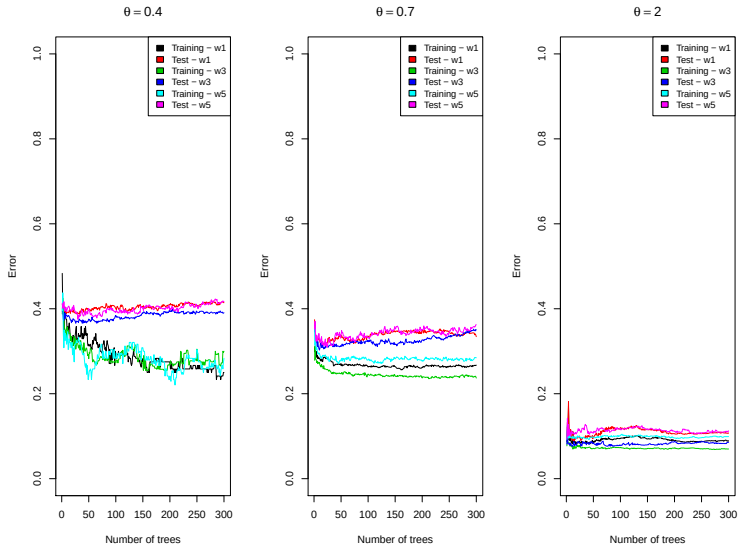


FIGURE 6.8: Boosting for nine simulated scenarios: different levels of homogeneity among the rankings, $\theta = (0.4, 0.7, 2)$, and three weights'structures, $w_1 = (1/3, 1/3, 1/3)$, $w_3 = (1/2, 1/2, 0)$ and $w_5 = (1, 0, 0)$

Real case application

Again, the real application is related to the dataset "vehicle" available in the UCI repository (*https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/vehicle/*). Figure 6.9 compares the outcomes obtained with the

weights' structures mentioned before: w_1, w_3 and w_5. The errors produced with w_1 and w_3 are quite similar, while w_5 penalizes the error as figured out with the simulations. The reason is clearer analyzing the variable importance: Figures 6.10 and 6.11 shows similar results and they reproduce the variable importance outcomes related to w_1 and w_3, respectively, while Figure 6.12 shows the different importance of the variables. Table 6.2 summarizes the outcomes of the variable importance measured using the three different vectors of position weights.

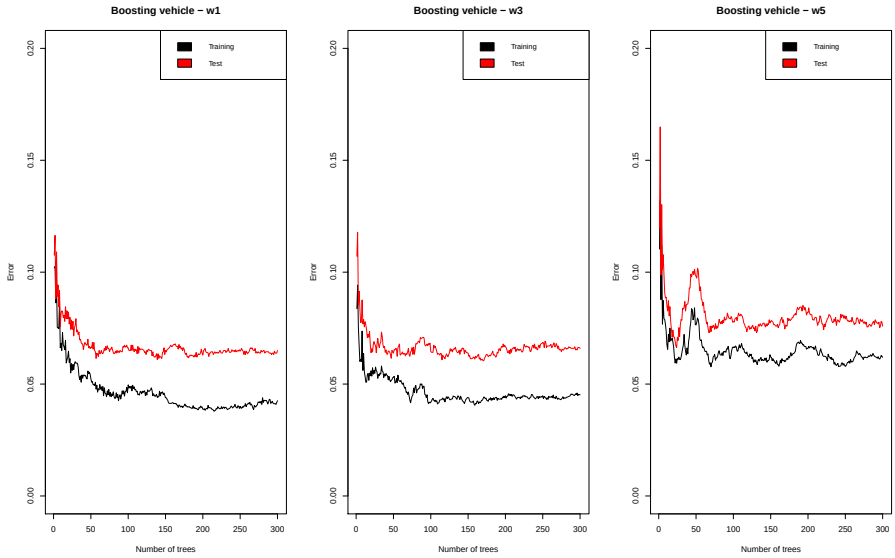


FIGURE 6.9: Training error and test error for AdaBoostM1 applied on the dataset "vehicle" using three different position weights: $w_1 = (1/3, 1/3, 1/3)$, $w_3 = (1/2, 1/2, 0)$ and $w_5 = (1, 0, 0)$.

6.3.4 Conclusion

In this section we propose Boosting for ranking data with different positions'weights structures. Looking at the single decision tree, we used the weighted Kemeny distance as a measure of impurity in the splitting process and its related weighted rank correlation coefficient (τ_x^w proposed by Plaia *et al.* (2019b) and Plaia *et al.* (2018a)) for identifying the median ranking in the final nodes. Once the trees are built up, τ_x^w was employed for assigning the median ranking as the final prediction, tree by tree, and for measuring the relative error. By means of simulations the sensitivity of the procedures to the different weighted structures was studied, in terms of error and variable importance. The considerations stated in Section 5.3, about future development of a faster "rpart" algorithm, is even more necessary since "adabag" need "rpart" objects. The extension to the definition and analysis of Random Forests for ranking data with position weights (and in future with item weights) will be necessary, for the sake of compeletness in the framework of Ensemble Methods for ranking data.

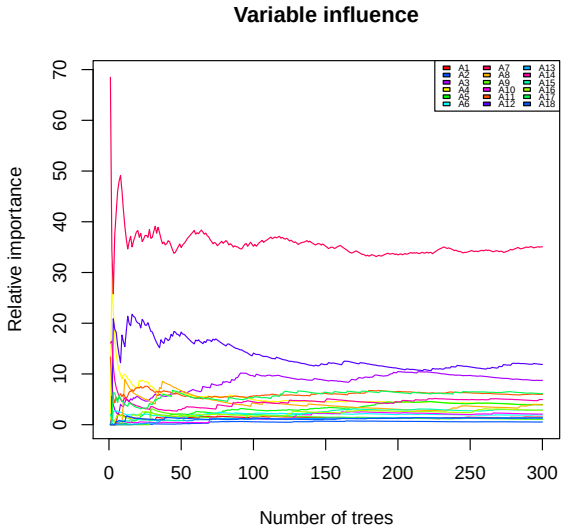


FIGURE 6.10: Variable importance for the dataset "vehicle" without positional weights for the rankings, i.e.: $w_1 = (1/3, 1/3, 1/3)$.

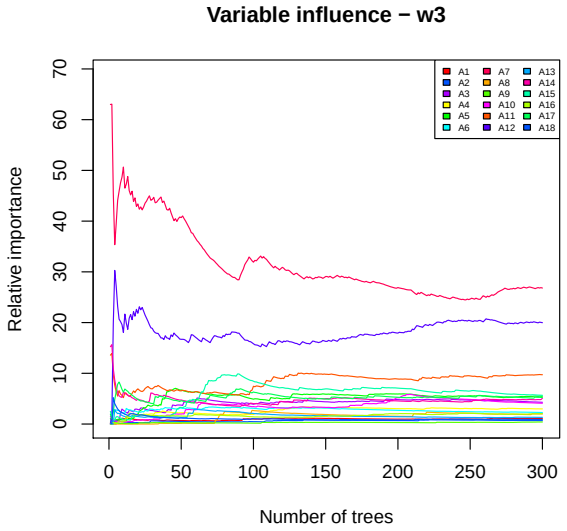


FIGURE 6.11: Variable importance for the dataset "vehicle" with positional weights given to the two first positions of the rankings: $w_3 = (1/2, 1/2, 0)$.

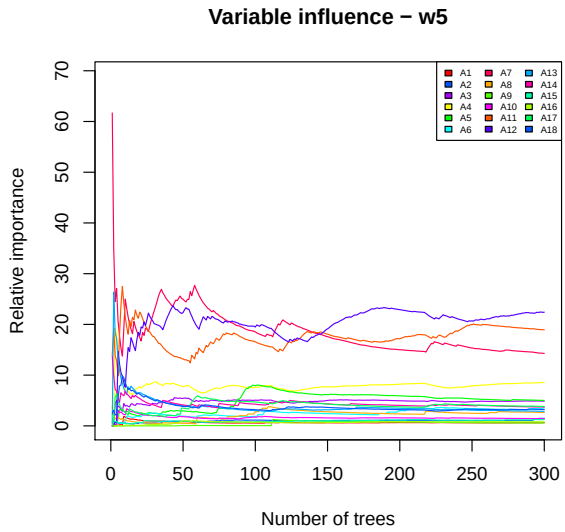


FIGURE 6.12: Variable importance for the dataset "vehicle" with positional weights giving importance only to the first position of the rankings: $w_5 = (1, 1, 0)$.

Variables	w_1	w_3	w_5
A1	1,49	1,11	0,63
A2	0,53	0,68	1,47
A3	8,73	4,29	4,88
A4	4	2,98	8,55
A5	3,96	5,47	5,03
A6	1,92	2,27	3,53
A7	35,04	26,81	14,29
A8	3,9	1,98	2,64
A9	1,04	0,4	0,56
A10	2,07	4,09	1,42
A11	5,97	9,72	18,94
A12	11,88	19,98	22,41
A13	1,35	1,32	2,86
A14	4,95	5,07	3,71
A15	2,88	5,66	1,24
A16	2,83	2,05	0,79
A17	6,17	5,23	3,82
A18	1,3	0,91	3,23

TABLE 6.2: Variable importance for the dataset "vehicle" with positional weights: $w_1 = (1/3, 1/3, 1/3)$, $w_3 = (1/2, 1/2, 0)$ and $w_5 = (1, 0, 0)$

Chapter 7

Model Selection in Linear Mixed-Effect Models: a review

7.1 Introduction

Linear Mixed effects Models (LMM) represent one of the most widely instruments for modelling data in applied statistics, and increasing research on linear mixed models has been rapidly in the last 10-15 years. This is due to the wide range of its applications to different types of data (clustered data such as repeated measures, longitudinal data, panel data, and small area estimation), which involve the fields of agriculture, economics, medicine, biology, sociology etc.

Some practical issues usually encountered in statistical analysis concern the choice of an appropriate model, estimating parameters of interest and measuring the order or dimension of a model. This paper focuses on model selection, which is essential for making valid inference. The principle of model selection or model evaluation is to choose the "best approximating" model within a class of competing models, characterized by a different number of parameters, a suitable model selection criterion given a data set (Bozdogan, 1987). The ideal selection procedure should lead to the "true" model, i.e. the unknown model behind the true process generating the observed data. In practice, one seeks, among a set of plausible candidate models, the parsimonious one that best approximates the "true" model.

The selection of only one model among a pool of candidate models is not a trivial issue in LMMs, and the different methods proposed in the literature over time are, often, not directly comparable. In fact, not only there is a different notation among papers and great confusion as regards the software (R, SAS, MATLAB, etc) to be used, but also a lack of landmarks allowing users to prefer one method rather than others.

Hence, the main purpose of this review is to provide a view about some useful components/factors characterizing each selection criterion, so that users can identify the method to apply in a specific situation. Moreover, we will also try to tidy up the notation used in the literature, by "translating", if necessary, the symbols and formulas found in each paper to produce a common "language". We begin by updating the recent review by Müller *et al.* (2013), then add some information about each selection criteria, such as the kind of effects that each method focuses on, or the structure of variance-covariance matrix, or the model dimensionality, or even the software used for implementing each method.

When coping with LMMs, it is not a good idea to assume independence or uncorrelation among response observations. For example, in the case of repeated measures: data are collected about the same individual over time. Hence, the traditional linear regression model is not appropriate to describe the data. For a detailed description of analogies and differences between linear mixed models and linear models, see (Müller *et al.*, 2013).

An important issue associated with LMMs selection is related to the dimension of the fixed and random components. Most of the literature bases inference, selection and interpretation of models in the finite (fixed) dimensional case, which means that the number of parameters is less than the number of units. Recently, more attention has been given to the handling of high-dimensional settings, which requires more complex computational applications. The word "high-dimensional" refers to situations where the number of unknown parameters that are to be estimated is one or several orders of magnitude larger than the number of samples in the data (Bühlmann and van de Geer, 2011). Furthermore, in LMMs, the number of parameters can grow exponentially with the sample size, i.e. the number of effects is strictly related to the number of units. Thus, if the sample size increases the set of effects diverges. Only recently some authors have tried to make inference within the LMM framework, on high-dimensional settings (Fan and Li, 2012; Schelldorfer *et al.*, 2011).

Model selection is a challenge in itself when one deals with the classic linear model. It becomes more complex when mixed models are involved, because of the presence of two kinds of effects with completely different characteristics and roles . Among others, a key aspect of linear mixed model selection is how to identify the real important random effects, i.e. those whose coefficients vary among subjects. It is important to note that the exclusion of relevant effects has a drawback on the estimation of the fixed effects: their variance-covariance matrix would be underfitted and the estimation of the variances related to the fixed part

estimates would be biased. The inclusion of irrelevant random effects in a model, on the other hand, would lead to a singular variance-covariance matrix of random effects, producing instability in the model (Ahn *et al.*, 2012). As pointed out by Müller *et al.* (2013), most procedures focuse on the selection of fixed effects exclusively. Only Chen and Dunson (2003) and Greven and Kneib (2010) worked on random part selection before Müller *et al.* (2013) . There are obvious difficulties due to computational issues in selecting only the random part, that is why the researchers who worked on the random effects, after Müller *et al.* (2013), optimise with respect to the fixed part, too, excepted for Li and Zhu (2013). In recent years, in fact, it has been easy to find procedures selecting both the effects .

It is worth noting that since the LMMs are a special case of Generalized LMMs, we obviously excluded from the current review all those methods built mainly for selecting effects in the GLMMs, such as Hui *et al.* (2017). Moreover, this review doesn't include works based on graphical tools for model selection if these graphical representations are refer to methods already existent in the literature. This is the case, for example, of Sciandra and Plaia (2018) who adapt an available graphical representation to the class of mixed models, in order to select the fixed effects conditioning on the random part and covariance structure, and of Singer *et al.* (2017) who discuss different diagnostic methods focusing on residual analysis but also addressing global and local influence, giving general guidelines for model selection.

This review mentions the available theoretical properties corresponding to the different methodologies, with comparisons among them whereas it's possible.

Müller *et al.* (2013) classified the proposed methods by considering four different kinds of procedures: information criteria (such as Akaike Information Criterion, Bayesian Information Criterion); shrinkage methods such as LASSO and adaptive LASSO; the Fence method and some Bayesian methods.

In this paper we prefer to cluster methods according to which part of the model, fixed, random or both, they focus on. The paper is organized as follows. In Section 2, we present the structure and notation of a linear mixed model and we discuss some problems occurring in selection models. In Section 3, we give an overview of model selection procedures within the LMMs framework that are useful for selecting linear mixed models, by considering the classification proposed in (Müller *et al.*, 2013). In Section 4 and 5, we describe the methods grouped according to the part of the model selected i.e. fixed and both, respectively. Finally, we

examine some simulations in Section 6, and conclude with a brief discussion and some conclusions in Section 7. Moreover, to help the reader decide which method to prefer, according to his own data, we include two tables (2 and 3), that summarize the main features of each method.

7.2 LMM and the Linear Mixed Model selection problem

Suppose data are collected from m independent groups of observations (called clusters or subjects in longitudinal data). The response variable $\boldsymbol{Y}_i$ is specified in the linear mixed models at cluster level as follows:

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, ..., m, \tag{7.1}$$

where $\boldsymbol{Y}_i$ is a n_i dimensional vector of observed responses, $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the known $n_i \times p$ and $n_i \times q$ matrices of covariates related to the fixed effects and to the random effects, respectively, $\boldsymbol{\beta}$ is the p-vector of unknown fixed effects, $\boldsymbol{b}_i$ is the q-vector of unobserved and independent random effects and $\boldsymbol{\epsilon}_i$ is the vector of unobserved random errors. Let us assume that $\boldsymbol{b}_i$s are independent of $\boldsymbol{\epsilon}_i$s and that they are independent and identically distributed random variables for each group of observations in the following way:

$$\boldsymbol{b}_i \sim N_q(0, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon}_i \sim N_{n_i}(0, \boldsymbol{\Sigma}), \tag{7.2}$$

where $\boldsymbol{\Psi}$ is a $q \times q$ positive definite matrix and $\boldsymbol{\Sigma}$ is a $n_i \times n_i$ positive definite matrix. Consequently, the response vector follows a multivariate normal distribution, $\boldsymbol{Y}_i \sim N_{n_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{V}_i)$, where the variance-covariance matrix is given by $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i^{'} + \boldsymbol{\Sigma}$.
The vectorized form of the model is:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}, \tag{7.3}$$

where all elements concern all macro units, therefore $\boldsymbol{Y}$ is a n dimensional vector ($n = \sum n_i$), $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the known $n \times p$ and $n \times q$ matrices of covariates related to the fixed effects and to the random effects, respectively, $\boldsymbol{\beta}$ is the p-vector of unknown fixed effects, $\boldsymbol{b}$ is the q-vector of unobserved and independent random effects and $\boldsymbol{\epsilon}$ represents the vector of unobserved random errors.

The selection of linear mixed effects models implies the selection of the "true" fixed parameters and/or the "true" random effects. Even if

there exists a kind of estimation for $\boldsymbol{b}$, the Best Linear Unbiased Predictors (BLUP, see Equation (7.7)), the correct investigation for identifying $\boldsymbol{b}$ requires to estimate its $q(q+1)/2$ variance-covariance parameters. Let $\boldsymbol{\tau}$ denote the s-vector filled with all distinctive components in the variance-covariance matrices $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$. A random effect is not relevant if its variance-covariance elements, for all observations, are zero (Ahn *et al.*, 2012), hence, it suffices to identify the non-zero diagonal components in $\boldsymbol{\Psi}$ (Wu *et al.*, 2016) correctly and, also, their related covariance terms, for avoiding the drawback of excluding random effects correlated to some explanatories.

We call $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ the overall set of parameters relevant in a linear mixed model. This set represents the whole group of the parameters related to the true model generating data. Let us identify the selection of linear mixed models with $M \in \mathcal{M}$, where $\mathcal{M}$ is the countable set containing all candidate models involved in the selection. The number of candidate models used depends on some contextual considerations: some variance-covariance components could be known or assumed to be known; some authors could focus only on nested models; or, still, the classic null model (the one with intercept only) could not be admitted among the set of candidate models (see Section 7.7 for further details).

The conditional log-likelihood for model (3) is given by:

$$l(\boldsymbol{\theta}|\boldsymbol{b};\boldsymbol{y}) = \log f_{\boldsymbol{y}}(\boldsymbol{y}|\boldsymbol{b};\boldsymbol{\theta}) =$$

$$-\frac{1}{2}\Big\{\log|\boldsymbol{\Sigma}| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{b})\Big\} - \frac{n}{2}\log(2\pi), \quad (7.4)$$

while the marginal likelihood is:

$$l(\boldsymbol{\theta};\boldsymbol{y},\boldsymbol{b}) = \log f_{\boldsymbol{y}}(\boldsymbol{y};\boldsymbol{b},\boldsymbol{\theta}) = -\frac{1}{2}\{\log|\boldsymbol{V}| + (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\}. \quad (7.5)$$

For fixed $\boldsymbol{\tau}$, the optimization process of the joint log-likelihood leads to an estimate of $\boldsymbol{\beta}$ that is similar to a generalized least squares estimator:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}) = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}. \quad (7.6)$$

The most popular approach for predicting $\boldsymbol{b}$ is an empirical Bayesian

method, which uses the posterior distribution $f(\boldsymbol{b}|\boldsymbol{y})$ yielding the following BLUP prediction:

$$\hat{\boldsymbol{b}}(\boldsymbol{\tau})_{BLUP} = \boldsymbol{\Psi}\boldsymbol{Z}'\boldsymbol{V}^{-1}\{\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})\}. \tag{7.7}$$

The same solutions of $\hat{\boldsymbol{\beta}}(\boldsymbol{\tau})$ and $\hat{\boldsymbol{b}}(\boldsymbol{\tau})_{BLUP}$ can be obtained by solving Henderson's linear mixed model equations (Müller *et al.*, 2013):

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{\Sigma}^{-1}\boldsymbol{Z} + \boldsymbol{\Psi}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}}(\boldsymbol{\tau}) \\ \hat{\boldsymbol{b}}(\boldsymbol{\tau}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}' \\ \boldsymbol{Z}' \end{bmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{y}. \tag{7.8}$$

Although consistent, the ML estimator of variance-covariance parameters is known to be biased in small samples. Hence, the restricted maximum likelihood estimators (REML) are used:

$$l_R(\boldsymbol{\tau}) = -\frac{1}{2}\{\log|\boldsymbol{V}| + \log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| + \boldsymbol{y}'\boldsymbol{P}^{-1}\boldsymbol{y}\}, \tag{7.9}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}$ (Müller *et al.*, 2013). Thus, the simple ML estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ will here forth be indicated as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\tau}}$, while the REML estimators as $\hat{\boldsymbol{\beta}}_R$ and $\hat{\boldsymbol{\tau}}_R$.

It is important to note that in many papers dealing with LMMs some authors use the σ^2 scaled versions of $\boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}$, which are $\sigma^2\boldsymbol{\Psi}_*$ and $\sigma^2\boldsymbol{\Sigma}_*$. Then we are going to use, in the description of the methods, the symbol $_*$ for those variance-covariance matrices scaled by σ^2.

7.3 Introduction to model selection criteria

Within the framework of linear mixed effect models, a large number of selection criteria are available in the literature. Model selection criteria are frequently set up by building estimators of discrepancy measures, which evaluate the distance between the "true" model and an approximating model fitted to the data.

7.3.1 AIC and its modifications

The most widely used criteria for model selection are the information criteria. Their application consists in finding the model that minimizes a function, in the form of a loss function plus a penalty, usually dependent on model complexity. The Akaike Information Criterion (AIC), introduced by Akaike (1992), is the most popular method. The Akaike Information Criterion is based on the Kullback-Leibler distance between

the true density of the distribution generating the data, $\boldsymbol{y}$, and, the approximating model for fitting the data, $g(\boldsymbol{\theta})$ (Vaida and Blanchard, 2005). With his criterion, Akaike tried to combine point estimation and hypothesis testing into a single measure, thus formalizing the concept of finding a good approximation of the true model in a predictive view. In this sense, a good model is the one that is able to generate predictive values (independent of the real data) as close as possible to the observed data. AI is given by $-2E_{f(\boldsymbol{y})}E_{f(\boldsymbol{y}^*)}\log g\{\boldsymbol{y}^*;\hat{\boldsymbol{\theta}}(\boldsymbol{y})\}$, where $\hat{\boldsymbol{\theta}}$ is an estimator of $\boldsymbol{\theta}$, while $\boldsymbol{y}^*$ represents the predictive set of data obtained from the fitted model and independent of $\boldsymbol{y}$. Vaida and Blanchard (2005) defined a new version of AI by conditioning the distribution $f(\boldsymbol{y};\boldsymbol{\theta})$ to the clusters. Hence, the conditional AI (cAI) uses the conditional distribution $f(\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{b})$ as follows:

$$-2E_{f(\boldsymbol{y},\boldsymbol{b})}E_{f(\boldsymbol{y}^*|\boldsymbol{b})}\log g\{\boldsymbol{y}^*;\hat{\boldsymbol{\theta}}(\boldsymbol{y}),\hat{\boldsymbol{b}}(\boldsymbol{y})\},$$

where $\hat{\boldsymbol{b}}(\boldsymbol{y})$ is the estimator of $\boldsymbol{b}$. It should be noted that $\boldsymbol{y}^*$ and $\boldsymbol{y}$ have to be considered conditionally independent of $\boldsymbol{b}$ and belonging to the same conditional distribution $f(.|\boldsymbol{b})$. These two last assumptions imply that they have the same random effects $\boldsymbol{b}$.

The underlying reasoning of the criterion based on the Akaike Information Criterion is not to identify the true model generating the data, but the best approximation of it, which adapts well to the data. The estimators employed for measuring AI and cAI are known as Akaike Information Criterion and conditional Akaike Information Criterion, respectively, and they are both biased for finite samples. They approximate their own information as minus twice the relative log-likelihood function plus a penalty term, $a_n(d_M)$, which tries to adjust the bias. The marginal AIC, defined by Vaida and Blanchard (2005), has the following generic formula:

$$\text{mAIC} = -2l(\hat{\boldsymbol{\theta}}) + 2a_n(p+q)$$

where $a_n = 1$ or $a_n = n/(n-p-q-1)$ in small samples (Sugiura, 1978; Vaida and Blanchard, 2005). The conditional Akaike Information Criterion (cAIC - Vaida and Blanchard (2005)) provides a procedure for selecting variables in LMMs with the purpose of predicting specific clusters or random effects, since the mAIC is inappropriate when the focus is on clusters and not on the population. For predicting at cluster level, the likelihood needs to be computed conditionally on the clusters and the random effects $\boldsymbol{b}_i$ need to be considered as parameters. Hence, for computing the cAIC, the terms to estimate are the $p+q+s$ parameters in $\boldsymbol{\theta}$. If all the variance elements $\boldsymbol{\tau}$ are known, the q random effects $\boldsymbol{b}$ are

predicted by the best linear unbiased predictor (BLUP), or using an estimated version of BLUP (Equation (7.7)). The generic formula for cAIC is:

$$\text{cAIC} = -2l(\hat{\boldsymbol{\theta}}|\hat{\boldsymbol{b}}) + 2a_n(\rho + 1) \tag{7.10}$$

where ρ is connected to the effective degrees of freedom used in estimating $\boldsymbol{\beta}$ and $\boldsymbol{b}$. Many authors ((Greven and Kneib, 2010; Kubokawa, 2011; Liang *et al.*, 2008; Shang and Cavanaugh, 2008; Srivastava and Kubokawa, 2010; Vaida and Blanchard, 2005)) have tried to reduce the bias of mAIC and cAIC, working on the penalty term in different ways, i.e. taking into account the MLE estimator or the REML estimator for $\boldsymbol{\theta}$, distinguishing if variance-covariance matrices are known or unknown. A clear and complete overview of all penalties used in the literature is available in (see Müller *et al.*, 2013, Secc 3.1 and 3.2).

7.3.2 Mallow's Cp

Another criterion, based on a discrepancy measure (Gauss discrepancy) and used for choosing the model nearest to the true one, is given by Mallows' C_p.

$$\text{C}_\text{p} = \frac{SSE_p}{\hat{\sigma}^2} - n + 2p,$$

with SSE_p and p representing, respectively, the error sum of squares and the number of parameters of the reference model and $\hat{\sigma}^2$ an estimate of σ^2 (Gilmour, 1996). Some variants on Mallows'C_p are provided by Kubokawa (2011) and are clearly presented by Müller *et al.* (2013).

7.3.3 BIC

The Bayesian Information Criterion is based on the marginal distribution of $\boldsymbol{y}$, which requires the full prior information about all parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ to be computed:

$$f(\boldsymbol{y}) = \int\int f_m(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\beta}, \boldsymbol{\theta})d\boldsymbol{\beta}d\boldsymbol{\theta}. \tag{7.11}$$

BIC, proposed by Schwarz (1978), is an approximation of $-2\log\{f_\pi(\boldsymbol{y})\}$, free of any prior distribution setup:

$$\text{BIC} = -2l(\hat{\boldsymbol{\theta}}) + (p + q)\log(N). \tag{7.12}$$

Since BIC is a Bayesian procedure for model selection, it requires prior distributions. Kubokawa and Srivastava (2010) derived the expression of

EBIC, an intermediate method between BIC and full Bayesian variable selection tools. The EBIC procedure employs partial non-subjective prior distribution only for the parameters of interest, ignoring the nuisance parameters in terms of distributional assumptions.

7.3.4 Shrinkage

Often, it is not feasible to compute Information Criteria in variable selection when p and/or q are large, i.e. in high-dimensional settings, when one deals with classic linear models. Hence, in this sense, shrinkage methods such as the least absolute shrinkage and selection operator, LASSO (Tibshirani, 1996), and its extensions such as the adaptive LASSO, ALASSO (Zou, 2006), the elastic net (Zou and Hastie, 2005) or the smooth clipped absolute deviation, SCAD (Fan and Li, 2012), have been proposed in the literature. When using these techniques, thanks to a penalization system, some coefficients are shrunk towards zero, while at the same time, the once influential on response are estimated to be non-zero. The shrinkage procedures are applicable to either the least squares or the likelihood functions. For the sake of simplicity, the penalized likelihood function is readopted in the case of the classical linear model:

$$-\sum_{i=1}^{n} l_i(\boldsymbol{\beta}; \boldsymbol{y}_i) + n\sum_{j=1}^{p} p_\lambda(||\boldsymbol{\beta}||_\ell), \tag{7.13}$$

where $||\boldsymbol{\beta}||_\ell$ is the ℓ-th norm of $\boldsymbol{\beta}$. Taking into account that ℓ_1 corresponds to work with the LASSO, while ℓ_2 refers to ridge estimation. The adaptive LASSO is an extension of LASSO. It involves the addition of some weights depending on the ℓ-th norm of $\boldsymbol{\beta}$, i.e. $p_\lambda(||\boldsymbol{\beta}||_\ell) = \lambda_j||\boldsymbol{\beta}||_\ell/2$, with $\lambda_j = \lambda/||\boldsymbol{\beta}||_\ell$, where ℓ is an additional parameter often considered equal to 1.

The generic SCAD penalty on $\boldsymbol{\theta}$ introduced by Fan and Li (2001) works on the first derivative of $p_\lambda(|\boldsymbol{\theta}|)$:

$$p'_{\lambda_j}(|\boldsymbol{\theta}|) = \lambda\left\{I(\boldsymbol{\theta} \leq \lambda) + \frac{(a\lambda - \boldsymbol{\theta})_+}{(a-1)\lambda} I(\boldsymbol{\theta} - \lambda)\right\}. \tag{7.14}$$

For the solution of $\boldsymbol{\theta}$, Fan and Li (2001) provided an algorithm via local quadratic approximations.

7.3.5 MDL principle

The minimum description length (MDL) principle originates from data compression literature and Rissanen (1986) who developed it to "understand" the observed data, it represents a valid statistical criterion employed for selecting linear mixed models. This method aims to detect the best model approximating the observed data, among a pool of candidate models, through a data compression process based on the code length needed to describe the data. A model can be described using fewer symbols than those necessary to describe the data. Usually, this criterion is used in the presence of independent data. Li *et al.* (2014) propose a MDL principle for fixed effects selection when there is a correlation between observations within clusters. The principle is presented as a good trade-off between AIC, thanks to its asymptotic optimality, and BIC, because of its consistency property. The proposed criterion is a hybrid form of MDL which merges a two stage description length and the mixture MDL with the dependent data.

7.4 Fixed effects selection

AIC and its modifications consist in finding the model that minimizes a function in the form of a loss function plus a penalty, which measures model complexity. Kawakubo and Kubokawa (2014) and Kawakubo *et al.* (2014) propose a modified conditional AIC and a conditional AIC under covariate shift in Small Area Estimation (SAE), respectively. For linear mixed model selection, random intercept model selection in particular, in the small area estimation, Marhuenda *et al.* (2013) work on two variants of AIC and two variants of the Kullback symmetric divergence criterion (KIC), defined as:

$$\text{KIC} = -2\log f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + 3(p+1).$$

Kawakubo and Kubokawa (2014) and Kawakubo *et al.* (2018) provide a modified version of the exact cAIC (McAIC), because the cAIC suggested by Vaida and Blanchard (2005) is highly biased when the candidate models do not include the true model generating the data (underspecified cases). They assume that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$, $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$, and extend cAIC to a procedure that could be valid both for the overspecified cases (situations in which the true model is included among the candidate models) and for the underspecified cases. The modified conditional

AIC is given by:

$$\text{McAIC} = -2\log f(y|\hat{\boldsymbol{b}}_j, \hat{\boldsymbol{\beta}}_j, \hat{\sigma}_j^2) + \widehat{\Delta_{cAI}}, \tag{7.15}$$

where $\widehat{\Delta_{cAI}}$ is the estimate of the bias of cAIC, estimated by:

$$\widehat{\Delta_{cAI}} = B^* + \widehat{B}_1 + \widehat{B}_2 + \widehat{B}_3, \tag{7.16}$$

where B^* is a function of $\boldsymbol{V}^{-1}$ and B_1, B_2 and B_3 are functions of $\boldsymbol{V}$ and $\boldsymbol{X}$. The authors demonstrate that B^*, $\widehat{B}_1$, $\widehat{B}_2$ and $\widehat{B}_3$ have distributions proportional to χ^2 with degrees of freedom opportunely quantified and, in the overspecified case, $\widehat{\Delta_{cAI}}$ reduces to B^*, i.e. McAIC=cAIC by Vaida and Blanchard (2005).

When the variable selection problem focuses on finding a set of significant variables for a good prediction, Kawakubo *et al.* (2014) propose a cAIC under covariate shift (CScAIC). They derive the cAIC of Vaida and Blanchard (2005) under the covariate shift for both known and unknown variances σ^2 and $\boldsymbol{\Psi}_*$ and with $\boldsymbol{\Sigma}_*$ assumed to be known.

The proposed criterion replaces, in the formula of the classic cAIC, the conditional density of $\boldsymbol{y}$ (the vector of the observed responses) given $\boldsymbol{b}$, with the conditional density of $\tilde{\boldsymbol{y}}$ (the vector of observed responses in the "predictive model": $\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{Z}}\boldsymbol{b} + \tilde{\boldsymbol{\epsilon}}$, a LMM with same regression coefficients $\boldsymbol{\beta}$ and random effects $\boldsymbol{b}$, but different -shifted- covariates) given $\boldsymbol{b}$.

$$\text{CScAIC} = -2\log g(\tilde{\boldsymbol{y}}|\hat{\boldsymbol{b}}, \hat{\boldsymbol{\beta}}, \hat{\sigma}^2) + B_c^*, \tag{7.17}$$

when σ^2 is unknown and estimated by its ML estimator and B_c^* is the bias correction.

Lombardía *et al.* (2017) introduce a mixed generalized Akaike information criterion, xGAIC, for SAE models. One typical model used in the field of SAE is the Fay-Herriot model, which is a particular type of LMMs containing only one random effect, the intercept. The clusters are represented by areas and the model in Equation (7.1) for each area is reduced to: $\boldsymbol{y}_i = \boldsymbol{x}_i^{'}\boldsymbol{\beta} + \boldsymbol{b}_i + \boldsymbol{\epsilon}_i$, with $i = 1, 2, ..., m$.
Instead of the usual AIC-types based only on the marginal or the conditional log-likelihood, the authors propose to use a new AIC, based on a combination of both the log-likelihood functions. The quasi-log-likelihood used for deriving the new statistics is the following:

$$\log(l_{\boldsymbol{x}}) = -\frac{1}{2}m\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{\mu})^{'}\boldsymbol{V}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}), \tag{7.18}$$

where $\boldsymbol{\mu} = E(\boldsymbol{Y}|\boldsymbol{b})$. The generalized degrees of freedom (xGDF), linked to the quasi-log-likelihood in Equation (7.18), takes into account the expectation and covariance with respect to the marginal distribution of $\boldsymbol{Y}$:

$$\text{xGDF} = \sum_{i=1}^{m} \frac{\partial E_{\boldsymbol{y}}(\hat{\boldsymbol{\mu}}_i)}{\partial(\boldsymbol{X}_i\boldsymbol{\beta})} = \sum_{i=1}^{m}\sum_{j=1}^{m} \boldsymbol{V}^{ij} cov(\hat{\boldsymbol{\mu}}_i, \boldsymbol{y}_j), \tag{7.19}$$

where $\boldsymbol{V}^{ij}$ is the ij-th element of the matrix $\boldsymbol{V}^{-1}$. Combining the $\log(l_{\boldsymbol{x}})$ with xGDF, the mixed generalized AIC is finally defined as:

$$\text{xGAIC} = -2\log(l_{\boldsymbol{x}}) + xGDF. \tag{7.20}$$

Han (2013) derives the closed form for the unbiased conditional AIC when the linear mixed model is reduced to the Fay-Herriot model. The author proposed a more suitable cAIC for three different approaches to fitting the model: the unbiased quadratic estimator (UQE), the REML estimator and the ML estimator. The unbiased cAIC for the Fay-Herriot model has the same form as for the classical LMMs, with i.i.d. errors (see Equation (7.10)), where the degrees of freedom are measured by $\Phi = \sum_{i=1}^{m} \frac{\partial \boldsymbol{X}_i\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{Y}_i} = tr(\frac{\partial \boldsymbol{X}'\hat{\boldsymbol{\beta}}}{\partial \boldsymbol{Y}})$, which is computationally expensive, because $\boldsymbol{X}_i\hat{\boldsymbol{\beta}}$ is not a linear estimator through $\hat{\sigma}_b^2$ and the derivates therein depend on the specific choise of estimating σ_b^2:

$$\text{cAIC} = -2\log f(\boldsymbol{y}|\boldsymbol{b}, \boldsymbol{\theta}) + 2\Phi. \tag{7.21}$$

If $\hat{\sigma}_b^2 = 0$, whatever is the method used for estimating it, then $\Phi = p$, otherwise when $\hat{\sigma}_b^2 > 0$ the way of measuring Φ is different. If the unbiased quadratic estimate method is used, then:

$$\Phi = \hat{\rho} + 2(m-p)^{-1} r' S \boldsymbol{\Sigma}^{-1} P^* r_s. \tag{7.22}$$

If $\hat{\sigma}_b^2 > 0$ is the REML or ML estimate:

$$\Phi = \hat{\rho} - 2\left(\frac{\partial \hat{s}}{\partial \boldsymbol{\sigma}_{\boldsymbol{b}}^2}\right)^{-1} r_s' \hat{\boldsymbol{\Sigma}}^{-1} P^* S \boldsymbol{\Sigma}^{-1} P^* r_s, \tag{7.23}$$

with $\frac{\partial \hat{s}}{\partial \sigma_b^2} = tr\big((\boldsymbol{\Sigma}^{-1}P^*)^2\big) - 2r_s'\hat{\boldsymbol{\Sigma}}^{-1}P^* r_s$ in the case of REML or $\frac{\partial \hat{s}}{\partial \sigma_b^2} = tr(\boldsymbol{\Sigma}^{-2}) - 2r_s'\hat{\boldsymbol{\Sigma}}^{-1}P^* r_s$ for ML estimating process, $P^* = I - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{\Sigma}^{-1}$, r the residuals from the OLS estimation for $\boldsymbol{\beta}$ and $r_s = \boldsymbol{\Sigma}^{-1}P^*\boldsymbol{Y}$ the standardized residuals obtained from the GLS estimation for $\boldsymbol{\beta}$. The closed-form cAIC results to be an unbiased estimator for the conditional AI for

the Fay-Herriot model.

It is worth mentioning Lahiri and Suntornchost (2015) for their contribution to the selection of fixed effects in LMMs with applications in SAE models, even if their proposal doesn't concern a modification of some Information Criteria. The authors define an alternative to the usual Mean Square Error and Mean Square Total, estimating them with $\widehat{MSE} = MSE - \overline{D}_w$ and $\widehat{MST} = MST - \overline{D}$, respectively, where $\overline{D}_w = \sum_{i=1}^{m}((1-h_{ii})D_i)/(m-p)$ and $\overline{D}_w = \sum_{i=1}^{m} D_i/m$, with $h_{ii} = \boldsymbol{x}_i^{'}(\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{x}_i$. They suggest to use $\widehat{MSE}$ and $\widehat{MST}$, because under standard regularity conditions these measures tend to the true MSE and MST with probability one, as the number of areas increases. But, since for small areas $\widehat{MSE}$ and $\widehat{MST}$ could be negative, the authors suggest an alternative to their estimates, through the function $h(x, b)$ in Equation (7.24) which guarantees to obtain positive values for them:

$$h(\boldsymbol{x}, \boldsymbol{b}) = \frac{2\boldsymbol{x}}{1 + \exp\left(\frac{2\boldsymbol{b}}{\boldsymbol{x}}\right)}. \tag{7.24}$$

This function allows to figure out new estimators in the following way: $\widehat{MSE} = h(MSE, \overline{D}_w)$ and $\widehat{MST} = h(MST, \overline{D})$.

Kubokawa and Srivastava (2010) derive an exact expression of the Empirical Bayes Information Criterion (EBIC) for selecting the fixed effects in a linear mixed model. Their criterion represents an intermediate solution between BIC and the full Bayes variable selection methods, because it exploits the partitioning of the vector of parameters $(\boldsymbol{\beta}, \boldsymbol{\tau}_*, \sigma)$ into two sub-vectors, one for the parameters of interest $(\boldsymbol{\beta})$ and the other one for the nuisance parameters $(\boldsymbol{\tau}_*, \sigma)$. Specifically, it works with a partial non-subjective prior distribution for only the parameters of interest, ignoring a prior setup for the nuisance parameters and applying the Laplace approximation for this one. The full prior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{\tau})$ can be written through a proper prior distribution, $\pi_1(\boldsymbol{\beta}|\boldsymbol{\tau}, \lambda)$, which is not completely subjective because of its dependence on an unknown hyperparameter λ:

$$\pi(\boldsymbol{\beta}, \boldsymbol{\tau}) = \pi_1(\boldsymbol{\beta}|\boldsymbol{\tau}, \lambda)\pi_2(\boldsymbol{\tau}).$$

The two authors derive EBIC, starting from the BIC but they approximate the marginal distribution of $\boldsymbol{y}$, $f(\boldsymbol{y})$, with one of its two components i.e. the conditional marginal density based on the partial prior distribution, $m_1(\boldsymbol{y}|\boldsymbol{\tau}, \lambda)$:

$$m_1(\boldsymbol{y}|\boldsymbol{\tau}, \lambda) = \int f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\tau})\pi_1(\boldsymbol{\beta}|\boldsymbol{\tau}, \lambda)d\boldsymbol{\beta}.$$

After estimating λ, $\hat{\lambda} = arg\max_{\lambda} m_1(\boldsymbol{y}|\hat{\boldsymbol{\tau}}, \lambda)$ using a consistent estimator of $\boldsymbol{\tau}$, the EBIC is obtained as follows:

$$\begin{aligned}\text{EBIC} &= -2log\{m_1(\boldsymbol{y}|\hat{\boldsymbol{\tau}}, \hat{\lambda})\} + dim(\boldsymbol{\theta})\log(n) \\ &= -2\log\{m_1(\boldsymbol{y}|\hat{\sigma}^2, \hat{\boldsymbol{\tau}}_*, \hat{\lambda})\} + (d+1)\log(n).\end{aligned}$$

The derivation of the EBIC neglects the full prior distribution, but it uses the non-subjective prior distribution $\pi_1(\boldsymbol{\beta}|\sigma^2, \lambda)$, assuming that, conditioned to σ^2, it assumes a multivariate normal distribution:

$$\pi_1(\boldsymbol{\beta}|\sigma^2, \lambda) = N_p(0, \sigma^2\lambda^{-1}W),$$

with an unknown scalar λ and a $p \times p$ known matrix W. A possible choice for W could be the so called Zellner's q-prior, $W_q = n(\boldsymbol{X}'\boldsymbol{X})^{-1}$. The authors demonstrate that EBIC is a consistent estimator.

Wenren and Shang (2016) and Wenren *et al.* (2016) work on conditional conceptual predictive statistics and on marginal conceptual predictive statistics for linear mixed model selection, respectively. The conditional C_p is formalized in both cases in which σ^2 and $\boldsymbol{\Psi}_*$ are known and unknown. The marginal C_p appears to be useful in two ways, both when the sample size is small and when there is a high correlation between the observations. Wenren *et al.* (2016) propose a modified variant of Mallows' C_p when there is a correlation between observations, even if not known. They work under the assumption that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$ and $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$. They assume that the estimator of the correlation matrix (for the candidate model) is consistent. The formalization of Modified C_p (MC_p) is as follows:

$$\text{MC}_\text{p} = \frac{SS_{RES}}{\hat{\sigma}^2} + 2p - n, \tag{7.25}$$

where SS_{RES} is the residual sum of squares for the candidate model, $\hat{\sigma}^2$ represents an asymptotically unbiased estimator for σ^2 and it is computed for the largest candidate model. MC_p is a biased estimator for the expectation of the transformed marginal Gauss discrepancy. However, it is an unbiased estimator of $\Delta_{C_p}(\boldsymbol{\theta})$, if the true model is included in the pool of all candidate models. For better performance, they also provide a more accurate estimator:

$$\text{IMC}_\text{p} = \frac{(n - p_* - 2)SS_{Res}}{SS^*_{Res}} + 2p - n + 2, \tag{7.26}$$

using the symbol * for referring to the largest candidate model. IMC_p results to be an asymptotically unbiased estimator of the expected overall

transformed Gauss discrepancy. It is preferred to MC_p because it avoids the bias introduced by $\frac{1}{\hat{\sigma}^2}$ used for estimating $\frac{1}{\sigma^2}$.

Wenren and Shang (2016) provide another conceptual predictive statistics for selecting a linear mixed model if one is interested in predicting specific clusters or random effects. Inspired by cAIC and conditional Mallow's C_p, they construct two versions of the conditional C_p (CC_p), according to known or unknown variance components. They work under the assumption that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$ and $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$, too. Assuming that σ^2 and $\boldsymbol{\Psi}_*$ are known, they combine a goodness of fit term with a penalty term, and propose CC_p defined as:

$$\text{CC}_\text{p} = \frac{SS_{Res}}{\sigma^2} + K, \tag{7.27}$$

where $K = 2\rho - n$ defines the effective degrees of freedom with $\rho = tr(H_1)$ (Hodges and Sargent, 2001). If the variance components are unknown, $\boldsymbol{\Psi}_*$ is substituted by its ML $\hat{\boldsymbol{\Psi}}_*$ or restricted MLE $\hat{\boldsymbol{\Psi}}_{*R}$ estimate. The effective degrees of freedom ρ is also estimated, $\hat{\rho} = tr(\hat{H}_1)$ where $\hat{H}_1 = \hat{H}_1(\hat{\boldsymbol{\Psi}}_*)$ or $\hat{H}_1 = \hat{H}_1(\hat{\boldsymbol{\Psi}}_{*R})$. σ^2 is estimated in the largest candidate model (*) through $\hat{\sigma}^2 = \frac{SS^*_{Res}}{N-p_*}$, an unbiased estimator of σ^2. For further details about $\hat{H}_1$ see Hodges and Sargent (2001). By substituting the variance components by their estimators in a suitable way, the conditional C_p is:

$$\text{CC}_\text{p} = (n - p_*)\frac{SS_{Res}}{SS^*_{Res}} + \hat{K}, \tag{7.28}$$

with $\hat{K} = 2\hat{\rho} - n$ indicating the (ML or REML) estimated penalty term.

Kuran and Özkale (2019) provide a conditional conceptual predictive statistic, too, in the framework of LMMs but applying a ridge estimator for overcoming multicollinearity problems. Like Wenren and Chang (2016), they work under the assumption that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$ and $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$. When we have to manage multicollinearity problems, usually we delete one or more variables related to the fixed effects, but this could cause some not irrelevant consequences: the fitted candidate model could be misspecified. For this reason, the two authors are motivated to require to the ridge estimator and the ridge predictor for LMMs proposed by Liu and Hu (2013) and Özkale and Can (2017):

$$\hat{\boldsymbol{\beta}}_k = (\boldsymbol{X}'\boldsymbol{V}_*^{-1}\boldsymbol{X} + kI_p)^{-1}\boldsymbol{X}'\boldsymbol{V}_*^{-1}\boldsymbol{y}, \tag{7.29}$$

$$\hat{\boldsymbol{b}}_k = \boldsymbol{\Psi}_*\boldsymbol{Z}'\boldsymbol{V}_*^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_k), \tag{7.30}$$

where k, a positive real number, represents the ridge biasing parameter. Its selection is obtained by minimizing a generalized cross-validation in the predictive step, while the same is measured through the minimization of the scalar mean square error of the ridge regression, in the estimation process (see Özkale and Can (2017)). Following Wenren and Shang (2016), they propose two versions of the conditional conceptual predictive statistic, distinguishing the case in which σ^2 and $\boldsymbol{\Psi}_*$ are known or they aren't. The proposed criteria are the same of CC_p in Equation (7.27) and in Equation (7.28), substituting the effective degrees of freedom under ridge estimator for LMMs, $\rho_k = tr(H_{1k})$, to ρ, $\hat{\rho}_k = tr(\hat{H}_{1k})$ to $\hat{\rho}$ and $SS_{Res,k} = (\boldsymbol{y} - \hat{\boldsymbol{y}}_k)'(\boldsymbol{y} - \hat{\boldsymbol{y}}_k)$ to SS_{Res}, where $H_{1k} = I_n - \boldsymbol{V}_*^{-1}[I_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}_*^{-1}\boldsymbol{X} + kI_p)^{-1}\boldsymbol{X}'\boldsymbol{V}_*^{-1}]$.

Li *et al.* (2014) proposed a two stage method based on the MDL principle. When $\boldsymbol{\beta}$ is the only unknown parameter, encoding the estimated parameter represents the first stage. Then, all the sequence of data with the distribution $f_{\hat{\boldsymbol{\theta}}}$ is encoded. The resulting total length code used for transmission is equivalent to BIC:

$$L(\boldsymbol{y}) = L(\boldsymbol{y}|\hat{\boldsymbol{\theta}}) + L(\hat{\boldsymbol{\theta}}) = -\log f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{y}) + \frac{p}{2}\log(m).$$

The penalty term, which measures the precision used to encode each parameter, is $\log(m)/2$ with a uniform distribution. The authors follow the idea of the mixture MDL proposed by Hansen and Yu (2003), which assumes a mixture distribution induced by the user-defined probability distribution $w(\boldsymbol{\theta})$ on the parameter space $\boldsymbol{\Theta}$. They assume that $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$, $\boldsymbol{\beta} \sim N(0, c\sigma^2(X_i'\boldsymbol{\Psi}_{*i}^{-1}\boldsymbol{X}_i)^{-1})$ and the hyperparameter c is a scalar constrained to be non negative. As regards the distribution of σ^2, an inverse gamma distribution is assumed with parameters $(a, 3/2)$. Hence, the mixture description length of $\boldsymbol{y}$ is expressed as:

$$-\log m(\boldsymbol{y}) = -\log \int f_{\boldsymbol{\theta}}(\boldsymbol{y}) w(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The code length is minimized with respect to $c \geq 0$ and the resulting $\hat{c}$ is plugged into the code length expression, leading to the $lMDL_0$ criterion. The expression of the final code length, with only $\boldsymbol{\beta}$ unknown and ignoring the impact of $\boldsymbol{b}$, is:

$$\begin{cases} \frac{1}{2}\left\{ \sum_{i=1}^{n} \boldsymbol{y}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i - FSS_\sigma + p\left[1 + \log\left(\frac{FSS_\sigma}{p}\right)\right] + \log n \right\}, & \text{if } FSS_\sigma > p, \\ \frac{1}{2}\sum_{i=1}^{n} \boldsymbol{y}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i, & \text{otherwise,} \end{cases}$$

$FSS_\sigma = (\sum_{i=1}^n \boldsymbol{y}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)(\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{X}_i)^{-1}(\sum_{i=1}^n \boldsymbol{X}_i'\boldsymbol{\Sigma}_i^{-1}\boldsymbol{y}_i)$ and $(\log n)/2$ represents the code length necessary for transmitting $\hat{c}$. If $FSS_\sigma \leq p$, $\hat{c} = 0$ and this implies that all fixed effects are null. The $lMDL_0$ criterion has the same structure of penalized likelihoods such as AIC and BIC, but with a proper data-adaptive penalty, depending on the covariance matrices. The two-stage mixture MDL principle, in the most realistic case with $(\sigma^2, \boldsymbol{\Psi}_*)$ unknown, it consists in estimating $\boldsymbol{\Psi}_*$ and plugging it into the code length. Minimization of the code length function, with respect to a and c, leads to an even more complex lMDL structure. The authors showed that the MDL criteria possess the selection consistency of BIC for finite-dimensional models.

Marino *et al.* (2017) give a really important contribution to the selection of relevant covariates in the LMMs, since their proposal is aimed at mixed models with missing data. Their work deals with selection of covariates in multilevel models, hence applicable to linear mixed models being a two-level model. The authors work under the assumption that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$ and $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$ and that parts of the covariates are ignorable missing, hence imputable. They propose to identify the covariates with missing data, to perform imputations producing m complete datasets (multiple imputations) and in the end to stack all these datasets into one single wide complete dataset. Before imputation, the generic linear mixed model in Equation (7.1) is rewritten, taking into account for the missing values, as follows:

$$\mathbf{Y}_i = \sum_{l=1}^{L}\sum_{g=1}^{G}(\mathbf{X}_{ig}^{(l)}\boldsymbol{\beta}_g^{(l)}) + \boldsymbol{Z}_i^{(\bullet)}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \quad i = 1, 2, ..., m;\ g = 1, ..., G;\ l = 1, ..., L; \tag{7.31}$$

where $\boldsymbol{X}_{ig}^{(l)}$ represents the g-th predictor for the i-th cluster from the l-th imputed dataset. After grouping all datasets into one, according to group relevant variables for imputation, the model could be rewritten in a compact way:

$$\mathbf{Y}_i = \mathbf{X}_i^{(\bullet)}\boldsymbol{\beta}^{(\bullet)} + \boldsymbol{Z}_i^{(\bullet)}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \tag{7.32}$$

where $\mathbf{X}_i^{(\bullet)} = (\mathbf{X}_{i1}^{(\bullet)}, \mathbf{X}_{i2}^{(\bullet)}, ..., \mathbf{X}_{iG}^{(\bullet)})'$ containing all the imputation data, and $\boldsymbol{\beta}^{(\bullet)}$ is the related G-vector of parameters. For identifying the relevant covariates, the authors suggest a shrinkage estimation process, i.e. to maximize the profile penalized REML log-likelihood built for the extended

model to imputed datasets:

$$Q_R(\boldsymbol{\beta}^{(\bullet)}) = l_R(\boldsymbol{\beta}^{(\bullet)}, \sigma^2, \boldsymbol{\Psi}_*) - \lambda \sum_{g=1}^{G} \sqrt{u_g} ||\beta_g^{(\bullet)}||, \tag{7.33}$$

where λ is the positive tuning parameter, u_g is the number of covariates, belonging to the group g, with imputation data inside. In case of no missing data or only one imputation, the optimal penalized solution is obtained through the classical LASSO penalization. Instead of maximizing the Equation (7.33), because of some computational issues, the authors prefer to solve a different optimization problem through an iterative algorithm concerning the following penalized function:

$$Q_R^2(\boldsymbol{\beta}^{(\bullet)}) = l_R(\boldsymbol{\beta}^{(\bullet)}, \sigma^2, \boldsymbol{\Psi}_*) - \sum_{g=1}^{G} \boldsymbol{\tau}_g^2 - \lambda^2 \sum_{g=1}^{G} \frac{u_g}{4\boldsymbol{\tau}_g^2} [||\beta_g^{(\bullet)}||]^2, \tag{7.34}$$

Hossain *et al.* (2018) propose a non-penalty Stein-like shrinkage estimator and then an adaptive version of the same estimator. This approach, first, consists in using a non-penalty Shrinkage Estimator (SE) and then it applies an adaptive measure related to the number of restrictions, which measures the distance between the restricted and the full model. The procedures works as follows: they propose to maximize the log-likelihood function under the postulated restricted parameter space, using the Lagrange multiplier vector, to get a restricted estimator for $\boldsymbol{\beta}$ this allows to build the profiling log-likelihood for estimating $\boldsymbol{\tau}$. Once the RE for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ are available, the likelihood ratio test statistic $D_m = 2[l(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}) - l(\hat{\boldsymbol{\theta}}_{RE}|\boldsymbol{\theta})]$ is introduced, and it allows to define the pretest estimator (PT) for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}_{\text{PT}} = \hat{\boldsymbol{\beta}} - I(D_m \leq \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{RE}). \tag{7.35}$$

Since that $\hat{\boldsymbol{\beta}}_{PT}$ is a discontinuous function of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{RE}$ and it depends on the α-level chosen a priori by the user, an adapted shrinkage estimator is built up, as follows:

$$\hat{\boldsymbol{\beta}}_{\text{PSE}} = \hat{\boldsymbol{\beta}}_{RE} + (1 - (r-2)D_m^{-1})(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{RE}), \quad r \geq 3, \tag{7.36}$$

The shrinkage estimator is, actually, a linear combination of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}_{RE}}$: $\lambda\hat{\boldsymbol{\beta}} + (1-\lambda)\hat{\boldsymbol{\beta}}_{RE}$, where the shrinkage parameter λ is an optimal value equal to $(r-2)D_m^{-1}$. The final estimator proposed by the authors is the

positive-part shrinkage estimator, which takes into account only the positive values of the estimator in Equation (7.36) due to the not convex function of SE in $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}_{RE}}$.

Only two papers discuss the selection of fixed effects in a linear mixed model in the case of a high dimensional setting: Rohart *et al.* (2014) and Ghosh and Thoresen (2018).

In many fields, it happens that one has to manage quite large amount of covariates. Thus, if interest is focused on obtaining an optimal inference, then, choosing only the relevant covariates is particularly important.

Ghosh and Thoresen (2018) contribute to linear mixed effects model selection with a non-concave penalization for the selection of fixed effects. Their procedure works with a maximum penalized likelihood, where non-concave penalties are implemented, considering $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$. A general objective function (with a general non-convex optimization):

$$Q_{n,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta}) = L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} P_{n,\lambda}(|\beta_j|), \tag{7.37}$$

has to be minimized with respect to $(\boldsymbol{\beta}, \boldsymbol{\eta})$ for a general loss function, $L(\boldsymbol{\beta}, \boldsymbol{\eta})$, which is assumed to be convex only in $\boldsymbol{\beta}$ and non-convex in $\boldsymbol{\eta}$. We can distinguish two situations: the number of fixed effects is less than the number of observations ($p < n$) and a high-dimensional set-up where p is of non-polynomial (NP) order of sample size n.

Making some appropriate assumptions on the penalty, it is important to note that: as n increases, $max\{p"_{\lambda_n}(|\boldsymbol{\beta}|)\} \to 0$ and $\frac{p'_{\lambda_n}(\boldsymbol{\theta})}{\lambda_n} > 0$. Moreover, the true parameter $\boldsymbol{\beta}_0$ is divided into two sub-vectors $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_0^{(1)'}, \boldsymbol{\beta}_0^{(2)'})'$, where $\boldsymbol{\beta}_0^{(2)}$ is a null vector. If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \to \infty$, as n increases, we can be sure that the local minimiser exists and satisfies that $\hat{\boldsymbol{\beta}}^{(2)}$ is equal to 0. Concerning the case of high-dimensionality, when p is of non-polynomial (NP) order of sample size, one should take into account the SCAD penalty for obtaining an estimator that is simultaneously consistent and satisfies the oracle property (Fan and Li, 2001) of variable selection optimality for any suitably chosen regularization sequence λ_n. Under some particular assumptions (extensively presented in Ghosh and Thoresen (2018)) what happens is that a local minimiser is obtained, which satisfies, with a probability of reaching one as n increases, that $\boldsymbol{\beta}^{(2)} = 0$ and that the estimated active set of $\hat{\boldsymbol{\beta}}$ coincides with the true active set of the fixed effect parameters. The $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\eta}}$ estimators are normally distributed under both types of dimensional settings.

Rohart *et al.* (2014) focus on the selection of the fixed effects in a high dimensional linear mixed model, suggesting the addition of an ℓ_1-penalization on $\boldsymbol{\beta}$ to the log-likelihood of the complete data. This penalization is useful in cases where the number of fixed effects is greater than the number of observations: it shrinks some coefficients to zero.They propose an iterative multicycle Expectation Conditional Maximization (ECM) algorithm to solve the minimization problem of the objective function:

$$g(\boldsymbol{\theta}; \boldsymbol{x}) = -2L(\boldsymbol{\theta}; \boldsymbol{x}) + \lambda||\boldsymbol{\beta}||_1, \tag{7.38}$$

The algorithm consists of four steps and it converges when three stopping criteria, based respectively on $||\boldsymbol{\beta}^{[t+1]} - \boldsymbol{\beta}^{[t]}||^2$, $||\boldsymbol{b}_k^{[t+1]} - \boldsymbol{b}_k^{[t]}||^2$ and $||L(\boldsymbol{\theta}^{[t+1]}, \boldsymbol{x}) - L(\boldsymbol{\theta}^{[t]}, \boldsymbol{x})||^2$, are fulfilled. Since the estimation of $\boldsymbol{\theta}$ is biased, a good choice would be to use the algorithm only for estimating the support of $\boldsymbol{\beta}$ and, after that, to estimate $\boldsymbol{\theta}$ using a classic mixed model estimation, based on the model that contains the only J relevant fixed effects: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_j + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$. The regularization parameter λ is tuned with the BIC,

$$\lambda_{\text{BIC}} = \min_{\lambda}\{\log|\boldsymbol{V}_\lambda| + (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)'V_\lambda^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda) + d_\lambda \log(n)\}, \tag{7.39}$$

where d_λ is the number of non-zero variance-covariance parameters plus the number of non-zero fixed effects coefficients. Substituting the LASSO method in the second step with any other variable selection method that optimizes a criterion, the algorithm becomes a multicycle ECM. All these considerations are valid assuming independence between the random effects, i.e. if there are q random effects corresponding to q grouping factors. As regards the selection of the random effects, it suffices to observe quite a small variance of a random effect to remove it at one step of the algorithm. The algorithm produces the same results and the same theoretical properties of the lmmLasso method (Schelldorfer *et al.*, 2011) when variances are known or they are assumed to be known, but it is much faster.

7.5 Random effects selection

Testing if random effects exist is equivalent to testing the hypothesis whether their variance/covariance matrix is made by zeros. Some authors, like Zhang *et al.* (2016), worked on the identification of the covariance structure of random effects, and others such as Wang (2016) provided some

characterizations of the response covariance matrix that cause model non-identifiability. The common perspective of these works lies in providing a preliminary analysis before the selection of the effects in a linear-mixed model, without providing a tool for testing the significance of random effects. Li and Zhu (2013), instead, introduced a test for evaluating the existence of random effects in semi-parametric mixed models for longitudinal data, proposing a projection method. The two authors created a test with two estimates for the error variance, one consistent under the null hypothesis and the other consistent under both the null and the alternative. The idea was to compare the two estimates under the alternative hypothesis, leading to reject the null one in case of large values of the test. But the test showed to be not stable and powerful, because of the projection matrix of $\boldsymbol{Z}$ variables onto the space spanned by the $\boldsymbol{X}$ variables. Hence, the two authors propose a similar, but more powerful test, in the LMMs framework but without projections. For developing the test, no assumptions are necessary for the random effects or the random errors. The test is built using the trace of the variance/covariance matrix of random effects:

$$T_{m\Omega} = \frac{tr(\hat{A})}{\sqrt{(\hat{k} - 3\hat{\sigma}^4)tr\{diag^2(M_{0m}^{tr})\} + 2\hat{\sigma}^4 tr\{(M_{0m}^{tr})^2\}}} \xrightarrow{d} N(0,1), \quad (7.40)$$

with m $\to \infty$. Under the alternative, the same test converges in distribution to $N(m_\Omega, 1)$, where

$$m_\Omega = \frac{k_0 c_{11} - q_1 + (q_1 - 1)c_{13} tr(\boldsymbol{\Sigma_z} Q_{10})}{\sqrt{(k - 3\sigma^4)C_{diag} + 2\sigma^4 C_{tr}}}, \quad (7.41)$$

with c_{11} and c_{13} estimates of variance/covariance matrices related to scaled $\boldsymbol{Z}$, C_{tr} and C_{diag} two non-negative constants such that $\lim_{m\to\infty}[m \cdot tr\{diag^2 (M_{0m}^{tr})\}] = C_{diag}$ and $\lim_{n\to\infty}[m \cdot tr(M_{0m}^{tr} 2)] = C_{tr}$. The test results to be consistent, not only under the null hypothesis, but under the alternative too. Even if the rate of convergence is slower than $m^{-1/2}$, the test is consistent. Furthermore, the test is good even if high correlations between $\boldsymbol{Z}$ and $\boldsymbol{X}$ are present.

7.6 Fixed and random effects selection

In most real cases, it is a matter of investigating the individuation of the important predictors corresponding not only to the fixed effects but, also, to the random part of the model. The joint selection of the two types of

effects has drawn more attention in recent years. Most of the proposed procedures are related to shrinkage methods: it suffices to look simultaneously at Table 7.2 and 7.3 to check this statement. The joint effect selection through penalized function can be based on a two-stage procedure, considering fixed and random effects separately, or a one-stage procedure, considering them jointly. Bondell *et al.* (2010) underlined that, in a separate selection, a change in the structure of one set of effects can lead to considerable different choices of variables for the other set of effects. Lin *et al.* (2013), on the other hand, argued that greater computation efficiency is reached if one prefer a separate selection of the effects. The number of stages employed in the shrinkage methods is reported in Table 6.1.

Braun *et al.* (2012) propose a predictive Cross-Validation (CV) criterion for the selection of covariates or random effects in the presence of linear mixed-effects models with serial correlation. Their approach is based on the logarithmic and the Continuous Ranked Probability Score (CRPS). Wang and Schaalje (2009) use point predictions, while Braun *et al.* (2012) focus on the whole predictive distribution, inspired by the proper scoring rules suggested by Gneiting and Raftery (2007), and the "mixed" cross-validation approach provided by Marshall and Spiegelhalter (2003). Going into detail, they use a very common proper score, the LS (local score), which considers the log predictive density $f(\boldsymbol{y})$ for the observed value $\boldsymbol{y}_{obs}$ and the CRPS, which is sensitive to the distance. The CRPS considers how close a predictive value is to the observed value through a ponderation system. With the univariate Gaussian as predictive distribution, the CRPS has the following form:

$$CRPS(\boldsymbol{Y}, \boldsymbol{y}_{obs}) = \sigma \left[\frac{1}{\sqrt{\pi}} - 2\varphi\left(\frac{\boldsymbol{y}_{obs} - \boldsymbol{\mu}}{\sigma} \right) + \right.$$
$$\left. - \frac{\boldsymbol{y}_{obs} - \boldsymbol{\mu}}{\sigma} \left(2\Phi\left(\frac{\boldsymbol{y}_{obs} - \boldsymbol{\mu}}{\sigma} \right) - 1 \right) \right], \tag{7.42}$$

where φ and Φ indicate the p.d.f. and the distribution function of a standardized Gaussian variable, respectively. The "mixed" cross-validation approach fits a model to the whole dataset. Once the hyperparameters have been estimated through all data, one observation is left out and for this one the LS and the CRPS are computed. Finally the cross-validation mean scores $\overline{LS}_{CV}$ and $\overline{CRPS}_{CV}$ are calculated from the distribution. The $\overline{LS}_{CV}$ is asymptotically equivalent to cAIC, but it is preferable to a full cross-validation approach because only one model is fitted at the beginning instead of fitting a model for each observation left out.

Schmidt and Smith (2016) focus on model selection when the number of models involved in the process is huge. They introduce a parameter subset selection algorithm (PSS). This technique consists in ranking the parameters by their significance, to establish the influential parameters. The basic assumption regarding the variance-covariance matrices of the random effects and of the random errors is $\boldsymbol{\Psi}$ and $\sigma^2 I_{n_i}$, respectively. The methodology is based on the asymptotic approximation of standard errors, measured through a normalization of the estimated standard deviations for each parameter. The proposed method works as follows: at first an estimate of the error variance is measured, then using a local sensitivity matrix - containing all the derivatives with respect to all fixed and random parameters for each i-th observation - one is able to estimate the variance-covariance matrix with all variances and correlations for the fixed and for the random effects (the authors suggest to use for instance the Moore-Penrose pseudoinverse). An estimate for the standard errors for each parameter is now possible: $\sqrt{Cov(k,k)}$, which is used for obtaining a measure of the selection score related to each k-th parameter in the i-th individual: $\alpha_{k_i} = |st.err._k / \hat{\boldsymbol{\theta}}_{k_i}|$. A small selection score is equivalent to a significant parameter. A ranking of all selection scores is created assigning a selection index γ_{k_i} according to the position reached by each α_{k_i} in the ordering. For all the parameters is calculated a global selection index $\Gamma_k = \sum_{i=1}^{m} \gamma_{k_i}$, which implies that the smallest values of this global index are related to the most significant parameters for all the clusters. If two or more parameters bring to the same Γ_k, then the parameter that has the smallest selection scores over all m individuals, is chosen as the most significant one. It is worth noting that since the PSS is repeated m times, the m sets of parameter rankings will be all different because the random effects parameter estimate will be different for each individual. The PSS algorithm attributes to the standard errors the role of measuring the parameter uncertainty: the parameters which obtain the smallest selection scores are those most significant and with the smallest uncertainty.

Rocha and Singer (2018) propose exploratory methods based on fitting standard regression models to the individual response profiles or to the rows of the sample within-units covariance matrix (in the case of balanced data) as supplementary tools for selecting a linear mixed-effects model. As concerns the choice of the fixed effects they examine the profile plots and suitable hypothesis tests. Assuming homoschedastic conditional independence, the model in eq (1) is re-written as:

$$\boldsymbol{y}_i = \boldsymbol{X}_i^* \boldsymbol{\beta}_i^* + \boldsymbol{\epsilon}_i, \tag{7.43}$$

where $\boldsymbol{X}_i^*$ contains the common variable between $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ and those that are unique to both the kind of variables, $\boldsymbol{\beta}_i^*$ contains the amount of $p+k$ parameters related to the fixed and the random effects. To test whether the generic $k-$th element of $\boldsymbol{\beta}$ is null, they propose the following statistic test:

$$t = \frac{\overline{\boldsymbol{\beta}}_k^*}{n^{-1}\sqrt{\hat{\sigma}^2 diag_k[(\sum_{i=1}^m \boldsymbol{X}_i^{*'}\boldsymbol{X}_i^*)^{-1}]}} \sim t_v, \tag{7.44}$$

where the degrees of freedom $v = \sum_{i=1}^m n_i - m(p+q)$ and the estimated $\hat{\sigma}^2$ is given by $\sum_{i=1}^m \frac{n_i-(p+q)}{v}\hat{\sigma}_i^2$, with:

$$\hat{\sigma}_i^2 = \frac{1}{n_i-(p+q)}\boldsymbol{Y}_i^{'}[I_{n_i} - \boldsymbol{X}_i^*(\boldsymbol{X}_i^*\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i]\boldsymbol{Y}_i. \tag{7.45}$$

The variance of $\hat{\beta}_{ik}^*$, $i = 1, 2, ..., m$, is expected to be equal to the $k-$th diagonal term of $\sigma^2(\boldsymbol{X}_i^{*'}\boldsymbol{X}_i^*)^{-1}$ when the variance of the corresponding random coefficient, $\hat{b}_{ik}$, is null. Otherwise, we might expect a larger variability of the $\hat{\beta}_{ik}^*$ around its mean. The $k-$th element of $\hat{\boldsymbol{\beta}}_i^*$, $\hat{\beta}_{ik}^*$, follows a $\mathcal{N}(\beta_{ik}^*; v_{ik}\sigma^2)$ distribution where $v_{ik} = diag_k\{(\boldsymbol{X}_i^{*'}\boldsymbol{X}_i^*)^{-1}\}$. Therefore, $\hat{\beta}_{ik}^*/\sqrt{v_{ik}} \sim \mathcal{N}(\beta_{ik}^*/\sqrt{v_{ik}}; \sigma^2)$. Letting $\hat{w}_{ik} = \beta_{ik}^*/\sqrt{v_{ik}}$ and $\overline{w}_k = \sum_{i=1}^m \hat{w}_{ik}/m$, it follows that:

$$t(\hat{w}_k) = \sqrt{n/(n-1)}(\hat{w}_{ik} - \overline{w}_k)/\hat{\sigma} \sim t_v. \tag{7.46}$$

Thus, for each k we expect around $\alpha\%$ of the values of $t(\hat{w}_k)$ outside the corresponding global significance level $\alpha^*\% = \alpha/(m(p+q))$ Bonferroni-corrected confidence interval, namely $[t_v(\alpha^*/2), t_v(1-\alpha^*/2)]$ where $t_v(\delta)$ denotes the $100\delta\%$ percentile of the t distribution with v degrees of freedom. A larger percentage of points outside that interval suggests that b_{ik} may be a random coefficient. Combining the two statistic tests in Equations (7.44) and (7.46) makes possible to detect which effects are statistically significant in the selection procedure. Another way to select the random effects requires the assumption of the homoschedastic conditional independence, i.e. when data are collected at the same time. In this case, the number of units for each i-th individual is the same and hence it's possible to estimate only one variance-covariance matrix $\boldsymbol{V}$ as $S - \hat{\sigma}^2 I_n$, where $S = (m-1)^{-1}\sum_{i=1}^m(\boldsymbol{y}_i - \overline{\boldsymbol{y}})(\boldsymbol{y}_i - \overline{\boldsymbol{y}})'$. Fitting polynomial models, with the same degree, to the rows of S the exploratory analysis along the lines obtained becomes an additional tool for the selection of the random effects.

7.6.1 One-stage shrinkage procedures

Chen *et al.* (2015) propose a variable selection methodology under the ANOVA type linear mixed models, for a high dimensional setting .They focus on the selection of the fixed effects and on testing the existence of the random effects. The authors state that $cov(\boldsymbol{b}_i) = \sigma_i^2 I_{n_i}$ and $\boldsymbol{\Sigma} = \boldsymbol{\sigma^2 I}$, without setting any distributional assumption for $\boldsymbol{Y}$. The selection regarding the fixed effects is made through the SCAD penalty. With the main purpose of removing the heteroschedasticity and correlation of the response variable, they modify the model in Equation (7.1), through an orthogonalization applied to random variables $\boldsymbol{Z}_\perp$. Let $\mathscr{M}(\boldsymbol{Z})$ be the vector space spanned by the columns of $\boldsymbol{Z}$, $\boldsymbol{Z}_\perp$ such that $\boldsymbol{Z}_\perp^{'}\boldsymbol{Z} = 0$, $\mathscr{M}(\boldsymbol{Z})^\perp$ the orthogonal complementary space of $\mathscr{M}(\boldsymbol{Z})$, therefore:

$$\boldsymbol{Z}_\perp \boldsymbol{Y} = \boldsymbol{Z}_\perp \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_\perp \boldsymbol{\epsilon}, \tag{7.47}$$

A sparse estimate of $\boldsymbol{\beta}$ can be obtained by minimizing:

$$Q(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{'} P_{(\boldsymbol{Z})_\perp}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + n\sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{7.48}$$

where $P_{(\boldsymbol{Z})_\perp} = \boldsymbol{Z}_\perp \boldsymbol{Z}_\perp^{'}$ is the orthogonal projection matrix of space $\mathscr{M}(\boldsymbol{Z})^\perp$ and $p_\lambda(\boldsymbol{\theta})$ is the SCAD penalty. Putting $\boldsymbol{Y}^* = \boldsymbol{Z}_\perp^{'}\boldsymbol{Y}$ and $\boldsymbol{X}^* = \boldsymbol{Z}_\perp^{'}\boldsymbol{X}$ the minimization algorithm $Q(\boldsymbol{\beta})$, the convergence test and the selection of thresholding parameters can be applied to Equation (7.48) without additional effort. Once the fixed effect parameters are estimated, the authors focus on the selection of the random effects, which means to detect if some $\sigma_i = 0$. The formal hypothesis system is:

$$H_0 : \sigma_k^2 = 0, k \in \mathscr{D} \leftrightarrow H_a : \exists \mathscr{D}_* \subseteq \mathscr{D}, s.t., \sigma_k^2 > 0, k \in \mathscr{D}_*, \tag{7.49}$$

where $\mathscr{D}$ is a subset of 1,2,...,q. Two estimators are proposed for σ^2: one, $\hat{\sigma}^2$, consistent even if the null hypothesis doesn't hold, the other one, $\hat{\sigma}_0$, consistent only under the null hypothesis. Indicating with $\hat{l} \hat{=} \{i : \hat{\beta}_i \neq 0\}$ all the relevant fixed effects, once the fixed parameters have been estimated, with $W_{\hat{l}} \hat{=} (\boldsymbol{X}_{\hat{l}}, \boldsymbol{Z})$ the relative covariate matrix together with the design matrix for the random effects, an estimate of σ^2 is defined as:

$$\hat{\sigma}^2 = \frac{\boldsymbol{Y}^{'} P_{(W_{\hat{l}})\perp} \boldsymbol{Y}}{tr[P_{(W_{\hat{l}})\perp}]}, \tag{7.50}$$

where $P_{(W_{\hat{l}})\perp}$ is the orthogonal projection matrix on the space of $\mathscr{M}(W_l)^{\perp}$:

$$\hat{\sigma}_0^2 = \frac{\boldsymbol{Y}' P_{(W_{\hat{l},-\mathscr{D}})\perp} \boldsymbol{Y}}{tr[P_{(W_{\hat{l},-\mathscr{D}})\perp}]}, \tag{7.51}$$

Let's assume that $\mathscr{D} = \mathscr{D}_1 \cup \mathscr{D}_2$ with $\mathscr{D}_1 \hat{=} \{k : k \in \mathscr{D}, m_k \to \infty \text{ when } n \to \infty\}$ and $\mathscr{D}_2 \hat{=} \{k : k \in \mathscr{D}, m_k = O(1)\}$. Under H_0 in (7.49), under certain conditions and assuming that the $\mathscr{D}_1$ is a null set, the authors built a test for assessing the existence of at least one of the random effects based on the difference between (7.50) and (7.51), which tends in distribution to $\chi^2(g)$ where g represents the dimension of space $\mathscr{M}(P_{(W_{l,-\mathscr{D}})\perp} Z_{\mathscr{D}})$. Whereas, under H_0 in (7.49) if $\mathscr{D}_1$ contains at least one element and knowing that $\hat{\sigma}^2 - \hat{\sigma}_0^2 = \boldsymbol{Y}' M_{n,\hat{l}} \boldsymbol{Y}$, with $M_{n,\hat{l}} \hat{=} \frac{P_{(W_{\hat{l}})\perp}}{tr(P_{(W_{\hat{l}})\perp})} - \frac{P_{(W_{\hat{l},-\mathscr{D}})\perp}}{tr(P_{(W_{\hat{l},-\mathscr{D}})\perp})}$, then the test to be considered is:

$$T_{nG,\hat{l}}(\gamma) = \frac{\boldsymbol{Y}' M_{n,\hat{l}} \boldsymbol{Y}}{\hat{\sigma}^2 \sqrt{\gamma tr\{diag^2(M_{n,\hat{l}})\} + 2tr\{M_{n,\hat{l}}\}}} \xrightarrow{d} N(0,1) \quad \text{as} \quad n \to \infty, \tag{7.52}$$

where γ indicates the kurtosis parameter that can be estimated with any consistent estimator.

Fan *et al.* (2014) propose a robust estimator for jointly selecting the fixed and random effects. The variable selection methodology defined by the three authors is robust against outliers in both the response and the covariates. The variance-covariance matrix of the random effects, is factorized using the Cholesky decomposition: $\boldsymbol{\Psi} = \Lambda\Gamma\Gamma'\Lambda$, where $\Lambda = diag(\nu_1, \nu_2, ..., \nu_q)$ and Γ represents a diagonal matrix and a triangular matrix with 1 on its diagonal, respectively. Hence, the random effects $\boldsymbol{b}_i$ are now substituted by $\Lambda\Gamma\boldsymbol{b}_i^*$. It's worth noting that setting to zero one element of Λ implies that all elements of the corresponding row and column in $\boldsymbol{\Psi}$ are zero, too, i.e. the relative random effect is not significant. To obtain a robust estimator which doesn't suffer the impact of outliers in the covariates, they introduce some weigths, w_{ij}, function of the Mahalanobis distance:

$$w_{ij} = min\left\{1, \left\{\frac{d_0}{(\boldsymbol{x}_{ij} - m_{\boldsymbol{x}})' S_{\boldsymbol{x}}^{-1} (\boldsymbol{x}_{ij} - m_{\boldsymbol{x}})}\right\}^{\frac{\delta}{2}}, \right.$$

$$\left. \left\{\frac{b_0}{(\boldsymbol{z}_{ij} - m_{\boldsymbol{z}})' S_{\boldsymbol{z}}^{-1} (\boldsymbol{z}_{ij} - m_{\boldsymbol{z}})}\right\}^{\frac{\delta}{2}}\right\}, \tag{7.53}$$

where the parameter $\delta \geq 1$, d_0 and b_0 are the 95-th percentiles of the chi-square distributions with the dimension of x_{ij} and z_{ij} like degrees of freedom, respectively. $S_{\boldsymbol{x}}$ and $S_{\boldsymbol{z}}$ are the median absolute deviance and $m_{\boldsymbol{x}}$ and $m_{\boldsymbol{z}}$ represent the medians of the covariates and random variables, respectively. For reducing the impact of outliers in the response variable, it is modified subtracting υ_{ij} to each its element in Equation (7.54), considering the studentized residuals $r_{ij} = y_{ij} - x_{ij}^{'}\beta - z_{ij}^{'}\Lambda\Gamma\boldsymbol{b}_i^*$

$$\upsilon_{ij} = sign(r_{ij})(|r_{ij}| - c)\sigma I(|r_{ij}| > c). \tag{7.54}$$

The robust log-likelihood is then defined as:

$$l^R(\boldsymbol{\theta}) = \log \int \sigma^{2^{-\frac{mq+n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left\|W^{\frac{1}{2}}(\boldsymbol{y}^* - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}I_m \otimes \Lambda I_m \otimes \Gamma \boldsymbol{b}^*)\right\|\right\} \times$$

$$\times \exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{b}^{*'}\boldsymbol{b}^*\right\}. \tag{7.55}$$

To guarantee the consistency property to the estimators, a correction has to be applied to $l_R(\boldsymbol{\theta})$:

$$l_C^R(\boldsymbol{\theta}) = l^R(\boldsymbol{\theta}) - a_m(\boldsymbol{\theta}), \tag{7.56}$$

with $a_m(\boldsymbol{\theta}) = \sum_{i=1}^m a_i(\boldsymbol{\theta})$ such that $\frac{\partial}{\partial \boldsymbol{\theta}} a_i(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left[\frac{\partial l_i^R(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$.

Selection and estimation of fixed and random effects are obtained maximizing:

$$Q^R(\boldsymbol{\theta}) = l_c^R(\boldsymbol{\theta}) - n(\sum_{j=1}^{p} p_{\lambda_n}(|\beta_j|) + \sum_{j=1}^{q} p_{\lambda m}(|\nu_j|)), \tag{7.57}$$

where $p_{\lambda m}(\cdot)$ is a shrinkage penalty with λ_n being the parameter which controls the amount of shrinkage, while $\overline{\boldsymbol{\beta}}_j$ and $\overline{\nu}_j$ are the unpenalized maximum estimators in Equation (7.55). The authors propose the ALASSO penalty to control the amount of shrinkage. For selecting λ_m the authors prefer to minimize the following BIC criterion:

$$\text{BIC}(\lambda) = -\frac{1}{2}\log|\hat{\boldsymbol{V}}| - \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2_{\hat{\boldsymbol{V}}} + \log(m)||\hat{\boldsymbol{\theta}}_\lambda||_0, \tag{7.58}$$

where $\hat{\sigma}^2$, part of $\hat{\boldsymbol{V}}$, is the median absolute deviation estimate, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{V}}$ are obtained as robust estimators and, finally, $||\hat{\boldsymbol{\theta}}_\lambda||_0$ states for the zero norm, measuring the amount of non-zero elements of $\hat{\boldsymbol{\theta}}_\lambda$.

Taylor *et al.* (2012) extend the two-parameter L_r penalty of Frank and Friedman (1993) and Fu (1998) in order to obtain new mixed model penalized likelihood, useful for selecting both the random and the fixed effects. The extended linear mixed model considers a set of penalized effects ($\boldsymbol{a}$), containing a subset of some effects:

$$\boldsymbol{y}|\boldsymbol{b} \sim N(\boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{Ma}, \boldsymbol{\Sigma}), \qquad \boldsymbol{y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Ma}, V(\boldsymbol{\tau})). \tag{7.59}$$

The authors use the scaled variance-covariance matrices $\boldsymbol{\Sigma}_* = \boldsymbol{\Sigma}/\sigma^2$ and $V(\boldsymbol{\tau})_* = V(\boldsymbol{\tau})/\sigma^2$ and identify $\boldsymbol{a}$, a potentially large vector of k effects, $k < p + s$ and $k < n$, with covariates $\boldsymbol{M}$. The penalized likelihood involves the L_r class of penalties with $0 < r < 1$:

$$l = \log f(\boldsymbol{y}, \boldsymbol{\theta}) - \sum_{j=1}^{k} p_\lambda(|\boldsymbol{a}_j|; r), \tag{7.60}$$

with the penalty term given by: $p_\lambda(|\boldsymbol{a}_j|; r) = \lambda((|\boldsymbol{a}_j| + 1)^r - 1)/r, \lambda > 0$. Taking into account a simple setting with $\sigma^2 = 1$ and $\boldsymbol{M}$ as orthonormal columns, an unbiased OLS estimator for $\boldsymbol{a}$ is obtained, through an iterative process:

$$\boldsymbol{a}_{j(s+1)} = sign(\hat{\boldsymbol{a}}_j)(|\boldsymbol{a}_j| - \lambda^*)_+. \tag{7.61}$$

This penalty is singular at origin, then, a local quadratic approximation is introduced to the derivative of the penalty, approximated as follows:

$$p_\lambda(|\boldsymbol{a}_j|; r) \approx \frac{1}{2}(\lambda(|\boldsymbol{a}_{js}| + 1)^{r-1}/|\boldsymbol{a}_{js}|)\boldsymbol{a}_j^2, \tag{7.62}$$

Thus, the introduction of a penalized term estimated iteratively, as shown is equivalent to inserting the pseudo-random effects in the linear mixed models. This it suffices to guarantee Henderson's results for estimation (REML estimates for $\boldsymbol{\tau}$) and prediction of both kinds of effects. Thresholding the elements of $|\boldsymbol{a}_{s+1}|$ with an optimal rule, a partitioned set of estimates into non-zero and zero components $(\boldsymbol{a}_{1,s+1}, \boldsymbol{a}_{2,s+1})$ is obtained. The zero set $(\boldsymbol{a}_{2,s+1}, \boldsymbol{M}_{2,s+1})$ is discarded from the set of information and the non-zero set replaces $\boldsymbol{a}_2$ until the iterative penalized REML estimates converge.

Li *et al.* (2018) propose a doubly regularized approach for selecting both the fixed and the random effects, in two cases: a) finite dimension of fixed and/or random effects, b) fixed and/or random effects that increase as the sample size goes to infinity. Their approach set $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$

and $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_* = \sigma^2 LL^{'}$, (Cholesky decomposition) with L a lower triangular matrix containing positive diagonal elements. The authors apply a double regularization (a ℓ_1-norm penalty for $\boldsymbol{\beta}$ and a ℓ_2-norm penalty for $\boldsymbol{\Psi}_*$ parameters) to the log-likelihood function, $l(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\Psi}_*)$ (equivalent to Equation (7.5)), as concerns the case with $m < p$. Hence, the objective function to maximize for estimating $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\Psi}_*$ is the following:

$$Q(\boldsymbol{\beta}, \boldsymbol{L}, \sigma^2) = \ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{L}) - \lambda_1 \sum_{j=1}^{p} |\beta_j| - \lambda_2 \sum_{k=2}^{q} \sqrt{L_{k1}^2 + ... + L_{kq}^2}. \quad (7.63)$$

For the case $m > p$, they modify $l(\cdot)$ in Equation (7.63) with the following function:

$$\ell_m(\boldsymbol{\beta}, \sigma^2, \boldsymbol{L}) = -\frac{1}{2}\sum_{i=1}^{m} \log|\sigma^2 \boldsymbol{V}_{*i}| - \frac{1}{2}\log\left|\sigma^{-2}\sum_{i=1}^{m} \boldsymbol{X}_i^{'}\boldsymbol{V}_{*i}^{-1}\boldsymbol{X}_i\right| +$$

$$-\frac{1}{2\sigma^2}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^{'}\boldsymbol{V}_{*i}^{-1}(\boldsymbol{Y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \quad (7.64)$$

The authors propose an algorithm as effective as the Newton-Raphson algorithm for estimating step by step $\boldsymbol{\beta}$ and L, since the penalty function in Equation (7.64) is separable.

Pan and Shang (2018b) propose a simultaneous selection procedure of fixed and random effects. Let's assume that $\boldsymbol{\Psi} = \sigma^2\boldsymbol{\Psi}_*$, $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$ and $\boldsymbol{\psi}$ containing the $\frac{q(q+1)}{2}$ unique elements in $\boldsymbol{\Psi}_*$, and let's indicate with $\boldsymbol{\theta}_*$ the vector related to $(\boldsymbol{\beta}, \boldsymbol{\psi})$. The authors maximize the following penalized profile likelihood function:

$$Q(\boldsymbol{\theta}_*) = p(\boldsymbol{\theta}_*) - \lambda_m \rho(|\boldsymbol{\theta}_*|) = \quad (7.65)$$

$$= -\frac{1}{2}\sum_{i=1}^{m} \log|\boldsymbol{V}_{i*}| - \frac{n}{2}\log\left(\sum_{i=1}^{m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^T\boldsymbol{V}_{i*}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right) - \lambda_m\rho(|\boldsymbol{\theta}_*|),$$

where λ_m is the tuning parameter controlling the amount of shrinkage and $\rho(|\boldsymbol{\theta}_*|$ is the adaptive Lasso function: $\rho(|\boldsymbol{\theta}_*| = |\boldsymbol{\theta}_*|/|\tilde{\boldsymbol{\theta}}_*|$, with $\tilde{\boldsymbol{\theta}}_*$ the MLE estimator of $\boldsymbol{\theta}_*$ used as the initial weights vector. To maximize 7.65, the authors use the Newton-Raphson algorithm, considering a local quadratic approximation at each iteration step as concerns the approximation of $|\boldsymbol{\theta}_*|$.

7.6.2 Two-stage shrinkage methods

One issue with the application of one stage shrinkage methods is that the combined dimension of both fixed and random effects is higher than the dimension of each of the two steps considered separately (Lin *et al.*, 2013). The computational efficiency depends also on the penalized log-likelihood taken into account for the selection of the random effects: the REML is preferred by Lin *et al.* (2013) and Pan (2016). The reasoning behind this choice is intuitive and underlined by Lin *et al.* (2013): REML estimators are unbiased and seem to be more robust to outliers than ML estimators. Furthermore, REML estimators do not involve the fixed effects.

Lin *et al.* (2013) propose two stage model selection by REML and pathwise coordinate optimization, inspired by the algorithm suggested by Friedman *et al.* (2007). The mixed model used is formulated assuming that $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$. In detail, during the first stage, the random effects are selected by maximizing the restricted log-likelihood penalized with the adaptive LASSO penalization:

$$Q^R(\boldsymbol{\tau}) = l^R(\boldsymbol{\tau}) - \lambda_{1,m} \sum_{j=1}^{s} \lambda_j w_j |\Psi_j|, \tag{7.66}$$

where Ψ_j is the j-th diagonal element of $\boldsymbol{\Psi}$ and w_j is the known weight. Because of the non-differentiable nature of the objective function, the Newton-Raphson algorithm is used for maximising $Q^R(\boldsymbol{\tau})$, after having locally approximated the penalty function by a quadratic function. Once the variance-covariance matrix is estimated, it is considered as known when the following penalized log-likelihood function is maximized to estimate the fixed effects:

$$Q^f(\boldsymbol{\beta}) = -\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})' \boldsymbol{v}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} w_j |\beta_j|. \tag{7.67}$$

Wu *et al.* (2016) propose an orthogonalization-based approach, which selects separately the fixed effects, at first, and then the random effects. All the selection steps are based on the least squares and no specific distribution assumption has to be involved. This method is suggested when the dimension of fixed effects is not large. The mixed model used considers $\boldsymbol{\Sigma} = \sigma^2 I$ and the selection procedure applies, at first, a QR decomposition of the design matrices, related to the random effects, for obtaining a homogeneous linear regression model (which does not depend on the

random effects). To select the fixed effects, it suffices to minimize, with respect to $\boldsymbol{\beta}$, the sum of residuals with SCAD penalization, thanks to possibility to find an unbiased estimate (Fan and Lì, 2001):

$$S_1(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' P_{\boldsymbol{z}'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})' + (n - ms)\sum_{j=1}^{p} p_{\lambda 1}(|\beta_j|), \tag{7.68}$$

where $P_{\boldsymbol{z}'} = I - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}$ is an idempotent matrix and $p_{\lambda 1}(|\boldsymbol{\beta}_j|)$ is a function whose first derivative depends on the tuning parameter λ. A ridge estimation process is computed for obtaining $\hat{\boldsymbol{\beta}}$, approximately:

$$\hat{\boldsymbol{\beta}}^{k+1} = (\boldsymbol{X}' P_{\boldsymbol{z}'}\boldsymbol{X} + (n - ms)\sum(\lambda_1, \hat{\boldsymbol{\beta}}^k))^{-1}\boldsymbol{X}' P_{\boldsymbol{z}'}\boldsymbol{Y}, \tag{7.69}$$

while to estimate σ^2 they consider:

$$W_2^*(\boldsymbol{\Psi}, \sigma^2) = \frac{1}{2}\sum_{i=1}^{m}((\boldsymbol{y}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) \otimes (\boldsymbol{y}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) - vec(\boldsymbol{V}_i))' \times \tag{7.70}$$

$$\times((\boldsymbol{y}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) \otimes (\boldsymbol{y}_i - \boldsymbol{x}_i\hat{\boldsymbol{\beta}}) - vec(\boldsymbol{V}_i)), \tag{7.71}$$

where $\boldsymbol{V}_i$ stands for the variance-covariance matrix of $\boldsymbol{Y}_i$, $\otimes$ for the Kronecker tensor product and $\hat{\boldsymbol{\beta}}$ for the estimates of the fixed effects obtained previously. Then, the objective function $S_2(\boldsymbol{\theta})$ with the SCAD penalty becomes:

$$S_2(\boldsymbol{\tau}) = \frac{1}{2}\sum_{i=1}^{m}(\tilde{\boldsymbol{Y}} - \boldsymbol{u}_i\boldsymbol{\tau})'(\hat{\boldsymbol{V}}_i \otimes \hat{\boldsymbol{V}}_i)^{-1}(\tilde{\boldsymbol{Y}} - \boldsymbol{u}_i\boldsymbol{\tau}) + \sum_{i=1}^{m} n_i^2 \sum_{j=1}^{(q^2+q)/2+1} p_{\lambda 2}(|\tau_j|), \tag{7.72}$$

and even in this situation it is solved iteratively obtaining the ridge estimation for $\boldsymbol{\tau}$:

$$\hat{\boldsymbol{\tau}}^{k+1} = (U'\hat{W}^{-k}U + \sum_{i=1}^{m} n_i^2 \sum_{\lambda 2}(\hat{\boldsymbol{\tau}}^k))^{-1}U'\hat{W}^{-k}\tilde{\boldsymbol{Y}}, \tag{7.73}$$

knowing that W is a diagonal matrix whose elements are given by $Wi = \boldsymbol{V}_i \otimes \boldsymbol{V}_i$, $\tilde{\boldsymbol{Y}}$ is the bias corrected $\boldsymbol{Y}$ and $\boldsymbol{u}_i$ is a function of $\boldsymbol{z}_i \otimes \boldsymbol{z}_i$.

Ahn *et al.* (2012) provide a class of robust thresholding and shrinkage procedures for selecting both the effects in linear mixed models. The robustness is guaranteed as they deal with non-normal correlated data and they do not assume any distribution of random effects and errors. For

the estimation of the variance components, a moment-based loss function is built. For ensuring the desired sparse structure they employ a hard thresholding estimator $\hat{\boldsymbol{\Psi}}^H = [\hat{\sigma}_{ij}^H]$, defined as $\hat{\sigma}_{ij}^H = \widetilde{\sigma}_{ij} I(|\widetilde{\sigma}_{ij}| > \nu)$, where $I(\cdot)$ is a typical indicator function and $\nu \geq 0$ is the parameter which controls the thresholding criterion. Although $\hat{\boldsymbol{\Psi}}^H$ is consistent, it could not be a positive semi-definite matrix in the presence of small sample sizes. Hence, in this sense, a sandwich estimator with a shrinkage penalty is yielded, by minimizing the following function:

$$Q_R(D) = \sum_{i=1}^{m} \sum_{j=1}^{n_i - 1} \sum_{k=j+1}^{n_i} (\widetilde{y}_{ijk} - z_{ij}' D \widetilde{\boldsymbol{\Psi}} D z_{jk})^2 + \lambda \sum_{i=1}^{q} d_i,$$
$$\text{subject to all } d_i \geq 0, \forall i = 1, ..., q.$$

To select the fixed effects, using $\boldsymbol{V} = \boldsymbol{Z}\widetilde{\boldsymbol{\Psi}}\boldsymbol{Z}' + \hat{\sigma}_\epsilon^2 I_n$, a Feasible Generalized Least Square (FGLS) estimator for $\boldsymbol{\beta}$ is computed as the minimiser of the following objective function:

$$Q_F(\boldsymbol{\beta}) = L_F(\boldsymbol{\beta} | \hat{\boldsymbol{\Psi}}, \hat{\sigma}_\epsilon^2) + \boldsymbol{\tau} \sum_{j=1}^{p} w_j |\beta_j|,$$

where data are transformed and w_j's are data-dependent weights.

Pan (2016) and Pan and Shang (2018a) propose a shrinkage method for selecting separately the two kinds of effects. The employment of the profile log-likelihood leads to a more efficient and stable computational procedure. Recalling the linear mixed model, let us assume that $\boldsymbol{\Psi} = \sigma^2 \boldsymbol{\Psi}_*$, $\boldsymbol{\Sigma} = \sigma^2 I_{n_i}$ and $\boldsymbol{\psi}$ contains the $\frac{q(q+1)}{2}$ unique elements in $\boldsymbol{\Psi}_*$. The profile and the restricted profile log-likelihood functions are, respectively:

$$p(\boldsymbol{\beta}, \boldsymbol{\psi}) = -\frac{1}{2} \sum_{i=1}^{m} \log |\boldsymbol{V}_i| - \frac{n}{2} \log \left(\sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) \right), \tag{7.74}$$

$$\begin{aligned} p_R(\boldsymbol{\psi}, \sigma) = &-\frac{1}{2} \log \left| \sum_{i=1}^{m} \boldsymbol{X}_i^T \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right| - \frac{1}{2} \sum_{i=1}^{m} \log |\boldsymbol{V}_i| \\ &- \frac{1}{2} (n - p) \log \left[\sum_{i=1}^{m} (\boldsymbol{y}_i - \boldsymbol{X}_i \tilde{\boldsymbol{\beta}})^T \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \tilde{\boldsymbol{\beta}}) \right], \end{aligned} \tag{7.75}$$

The random covariance structure is selected by maximizing the penalised restricted profile log-likelihood with the adaptive LASSO, but a factorization of the vector containing the variance-covariance elements of $\boldsymbol{\Psi}_*$ in $(\boldsymbol{d},\ \boldsymbol{\gamma})$ has to be carried out before hand, with $\boldsymbol{d}$ representing the vector of the diagonal elements and $\boldsymbol{\gamma}$ the vector of parameters that can vary freely:

$$Q_R(\boldsymbol{\psi}) = p_R(\boldsymbol{\psi}) - \lambda_{1m} \sum_{j=1}^{q} w_{1j} d_j|, \tag{7.76}$$

where λ_{1m} is the tuning parameter and $w_1 = 1/|\tilde{\boldsymbol{d}}|$ are weights used for reaching the optimality of the solution, with $\tilde{\boldsymbol{d}}$ computed as a root-n consistent estimator vector of $\boldsymbol{d}$. The Newton-Raphson algorithm is first applied for maximizing the penalized restricted profile likelihood function leading to $\hat{\boldsymbol{V}}$ and, then, the same is applied for maximizing the penalized profile likelihood function:

$$Q_F(\boldsymbol{\beta}) = p_F(\boldsymbol{\beta}) - \lambda_{2m} \sum_{j=1}^{p} w_{2j} |\beta_j|, \tag{7.77}$$

where $p_F(\boldsymbol{\beta})$ is the profile log-likelihood, λ_{2m} is the tuning parameter for fixed effect selection and w_{2j} are weights computed as the inverse of $|\tilde{\boldsymbol{\beta}}_j|$, considering that $\tilde{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. When the algorithm converges, the maximizer of the penalized profile log-likelihood is obtained. Hence, the set of suitable covariates is identified.

Fan and Li (2001) stated that "the penalty functions have to be singular at the origin to produce sparse solutions (many estimated coefficients are zero), to satisfy certain conditions to produce continuous models (for stability of model selection), and to be bounded by a constant to produce nearly unbiased estimates for large coefficients". The estimator obtained through the penalty functions should lead to three important properties: asymptotic unbiasedness for avoiding modeling bias; sparsity, i.e. as a thresholding rule, the estimator should shrink some estimated coefficients to zero in order to reduce model complexity; continuity in data to avoid instability in model prediction. They showed, in few, that the choice of the shrinkage parameter should guarantee the well known oracle properties in the resulting estimator: the penalized likelihood estimator is root-n consistent if $\lambda_n \to 0$, a set of estimated parameters is set to 0 and the remaining estimators converge asymptotically to a normal distribution when $\sqrt{n}\lambda_n \to \infty$.

Hossain *et al.* (2018) show that under certain regularity conditions and for fixed alternatives $B_{H_a} = \delta \neq 0$, as n increases, the estimators

$\hat{\boldsymbol{\beta}}_{PT}$ (see in Equation 7.35), $\hat{\boldsymbol{\beta}}_{PSE}$ (see in Equation 7.36) and the positive-part shrinkage estimator converge in probability to $\hat{\boldsymbol{\beta}}$ and they derive the asymptotic joint normality for the unrestricted and restricted estimators, of which the three estimators are a function. Fan *et al.* (2014) demonstrate that their proposed robust estimator enjoy all the properties defined by Liski and Liski (2008). Chen *et al.* (2015) demonstrate only the validity of the Oracle property of only sparsity and consistency, but not the asymptotical distribution. Li *et al.* (2018) show the "sparsistency" property which ensures the selection consistency for the true signals of both fixed and random effects, hence, they provide analytical proofs about the validity of consistency and sparsity, but nothing about the distributional form. Pan and Shang (2018b) demonstrate that their procedure fills the consistency and the sparsity properties, without mentioning anything about the asymptotical normality. Marino *et al.* (2017) only refer to take a look at Rubin (2004) in which is possible to assess that "a small number of imputations can lead to high-quality inference". As concerns Rohart *et al.* (2014) thus no mention about asymptotic properties fulfilled by their final estimator. Pan (2016), Pan and Shang (2018a), Ahn *et al.* (2012) and Lin *et al.* (2013) demonstrate that, if $\lambda \to 0$ and $\sqrt{m}\lambda \to \infty$ as $m \to \infty$, the estimators produced by their two stage model selection are $\sqrt{m}$ consistent and they possess the oracle properties, i.e. sparsity and asymptotic normality (asymptotically the proposed approaches can discover the subset of significant predictors). In other words, for an oracle procedure, the covariates with nonzero coefficients will be identified with probability tending to one, and the estimates of nonzero coefficients have the same asymptotic distribution as the true model (Pan, 2016). All these statements are valid if an appropriate tuning parameter is chosen.

Consistent variable selection depends on the choice of the tuning parameter. The shrinkage procedures yield estimates, assuming the tuning parameters as known, but they are not. Hence, they have to be tuned among a pool of values, from the largest to the smallest quantity, identifying a path through the model space. After constructing the path and reducing parameter space, one can apply a direct approach (information criteria, cross validation and so forth) to better identify the important variables. For this reason, shrinkage methods are, usually, employed in the case of many variables, thanks to the fact that they do not need to focus on all possible models (2^{p+q}). The most widely used methods in the literature for tuning the parameter, which controls regularization, are cross-validation and BIC. "A more rigorous theoretical argument justifying the use of the BIC criterion for the ℓ_1 penalized MLE in high-dimensional linear mixed effects models is missing: the BIC has been

empirically found to perform reasonably well"(Schelldorfer *et al.*, 2011). This seems to be generally valid for other shrinkage methods: there is not theoretical justification for employing the BIC. Fan *et al.* (2014) highlight their choice to select the shrinkage parameter through the BIC criterion is due to the fact that GCV leads to over-fitting models and AIC seems not to be consistent when the true model has a sparsity structure. The BIC criterion on which the authors base their selection of λ_n is the following:

$$\text{BIC}(\lambda) = -\frac{1}{2}\log|\hat{\boldsymbol{V}}| - \frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}||^2_{\hat{\boldsymbol{V}}} + \log(m)||\hat{\boldsymbol{\theta}}_\lambda||_0, \tag{7.78}$$

where $\hat{\boldsymbol{V}} = diag(\hat{\boldsymbol{V}}_1, \hat{\boldsymbol{V}}_2, ..., \hat{\boldsymbol{V}}_m)$ and the generic $\hat{\boldsymbol{V}}_i$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Psi}}_*$ are the robust estimates contained in $\hat{\boldsymbol{\theta}}_\lambda$ upon convergence of the EM algorithm. Because of the over-fitting problems using GCV, Marino *et al.* (2017) choose the BIC criterion for the selection of the tuning parameter:

$$\text{BIC}(\lambda) = -2l_R(\boldsymbol{\beta}^{(\bullet)}, \hat{\sigma}^2, \hat{\boldsymbol{\Psi}}_*) + q \times ln(n), \tag{7.79}$$

where $l_R(\boldsymbol{\beta}^{(\bullet)}, \hat{\sigma}^2, \hat{\boldsymbol{\Psi}}_*)$ is the REML log-likelihood function related to the model in (7.32).

Li *et al.* (2018) select the two tuning parameter minimizing a variant of BIC, proposed by Wang (2016):

$$\text{BIC} = -2p_R(\boldsymbol{\beta}, L) + \left[d_{\boldsymbol{\beta}} + \frac{(1 + d_{\boldsymbol{\Psi}_*})d_{\boldsymbol{\Psi}_*}}{2}\right]\log(n), \tag{7.80}$$

where $p_R(\boldsymbol{\beta}, L)$ is the profile log-likelihood in Equation (7.75), $d_{\boldsymbol{\beta}}$ and $d_{\boldsymbol{\Psi}_*}$ are given by the amount of non-zero elements in $\boldsymbol{\beta}$ and on the diagonal of $\boldsymbol{\Psi}_*$, respectively. Pan (2016) and Pan and Shang (2018a) propose to minimize the BIC or the AIC or the Generalized CV (GCV) as possible criteria for selecting the optimal tuning parameter. The above criteria, surely, have to be computed with the corresponding profile likelihood, shown in Equations (7.74) and (7.75), to identify the tuning parameter for the fixed part and the random part, respectively. The degrees of freedom necessary to compute all three criteria also refer to the fixed effects in one case (the number of non zero $\hat{\boldsymbol{\beta}}$'s) and to the random part in the other case (the amount of nonzero parts in $\hat{\boldsymbol{\psi}}$). Pan and Shang (2018b) select the optimal λ by minimizing the BIC criterion, where the degrees of freedom takes into account the number of non-zero elements in $\boldsymbol{\theta}_*$. The tuned parameters (λ_1, λ_2) are computed, by Wu *et al.* (2016), with a CV or GCV technique. Taylor *et al.* (2012) and Ahn *et al.* (2012) choose a tuning parameter that minimizes the BIC criterion, Taylor *et al.* (2012)

focus on the value of r (from a fixed grid, see Equation (7.60)), which leads to the minimum BIC, after obtaining convergence for the penalized REML estimators:

$$\text{BIC} = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{a}}, \hat{\boldsymbol{\tau}}) + \log(m)\#df, \tag{7.81}$$

where $l(\cdot)$ is the un-penalized (since it involves $\boldsymbol{a}$ as fixed effects) marginal loglikelihood over the random effects $\boldsymbol{b}$ evaluated at the REML estimates of $\boldsymbol{\tau}$ and $\#df$ represents the number of nonzero elements in $\hat{\boldsymbol{a}}$. Ahn *et al.* (2012) work on a modified version of the BIC, similar to the RSS ratio, for both the fixed effects and the random effects:

$$\text{BIC}_R(\nu) = \frac{L_0(\boldsymbol{\Psi}_\nu^H)}{L_0(\boldsymbol{\Psi})} + \frac{\log(n)}{n} \times df1, \tag{7.82}$$

$$\text{BIC}_F(\boldsymbol{\tau}) = \frac{L_F(\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}|\hat{\boldsymbol{\Psi}}, \hat{\sigma}^2)}{L_F(\hat{\boldsymbol{\beta}}_G|\hat{\boldsymbol{\Psi}}, \hat{\sigma}^2)} + \frac{\log(n)}{n} \times df2, \tag{7.83}$$

where $\hat{\boldsymbol{\beta}}_G$ is the FGLS estimator and $df1$ and $df2$ represent the number of nonzero components on the diagonal in $\hat{\boldsymbol{\Psi}}^H$ and in $\hat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$. The degrees of freedom measure the effective model dimension. Unlike Bondell *et al.* (2010) and Ibrahim *et al.* (2011), where the degrees of freedom considered are, respectively, sample size n and cluster size m, in the methods discussed above the number of parameters that can vary freely is connected to the nonzero parameters in the working model (fixed components and variance-covariance elements of the random effects). As pointed out by Müller *et al.* (2013), the number of nonzero estimated components related to the tuning parameter is not equivalent to the number of independent parameters, which is instead true for the linear models.
The main characteristics associated with shrinkage procedures available in the literature, are summarised in Table 6.1.

7.7 Review of simulations

Almost all the authors have performed at least one simulation to measure and demonstrate the reliability of their own procedure. As in a Meta-analysis, we have collected the simulations but, since the results are not directly comparable, the tables synthesise the main parameters characterizing the simulations. We followed the setting of Müller *et al.* (2013), for continuity to purposes. Considering Table 7.2, the smaller the values of $min|\boldsymbol{\beta}|/\sigma$ and $min\{ev(\boldsymbol{\Psi}/\sigma^2)\}$ the more difficult the selection of

the true model for β and τ. Nevertheless, it is worth noting that these values are not useful as regards the goodness of fit of the models or the real ability of the methods, once they are applied, for identifying the true values of β and τ, since they refer to initial settings of simulations and not to their results. As Müller *et al.* (2013) underlined, one could consider these simulations as a mere meta-analysis. The results obtained are not directly comparable, because the authors use different measures to assess the performance of their method.

Reference	Consistency	Sparsity	Asymptotic Normality	Number of λ_s	Selection of λ_s	Number of stages	Penalty
BKG10 (Bondell *et al.*, 2010)	✓	✓	✓	1	BIC	1	ALASSO
IZGG11 (Ibrahim *et al.*, 2011)	✓	✓	✓	2	BIC	1	SCAD, ALASSO
PL12 (Peng and Lu, 2012)	✓	✓	✓	2	GCV AIC BIC	2	SCAD
Inserted in Müller *et al.* (2013)							
AZL12 (Ahn *et al.*, 2012)	✓	✓	✓	2	BIC	2	Hard, Sandwich
CLSZ15* (Chen *et al.*, 2015)	✓	✓		1	GCV	1	orthogonality-based SCAD
FQZ14 (Fan *et al.*, 2014)	✓	✓	✓	1	BIC	1	ALASSO
GT16* (Ghosh and Thoresen, 2018)	✓	✓	✓	1	BIC	1	SCAD
HTA18* (Hossain *et al.*, 2018)	✓		✓	1	$(r-2)LRT^{-1}$	1	James-Stein
LPJ13 (Lin *et al.*, 2013)	✓	✓	✓	2	BIC	2	ALASSO
LWSWZZ18 (Li *et al.*, 2018)	✓	✓		2	BIC	1	LASSO:ℓ_1,ℓ_2
MBL17 (Marino *et al.*, 2017)	✓	✓		1	BIC	1	LASSO:ℓ_1,ℓ_2
P16 (Pan, 2016)	✓	✓	✓	2	BIC_R BIC_F GCV_R GCV_F AIC_R AIC_F	2	ALASSO
PS18 (Pan and Shang, 2018b)	✓	✓		2	BIC	1	ALASSO
RSL14 (Rohart *et al.*, 2014)				1	BIC	1	LASSO
TVCN12 (Taylor *et al.*, 2012)	✓	✓	✓	1	BIC_R	1	$L_R\ r<1$
WLXZ16 (Wu *et al.*, 2016)	✓	✓	✓	2	GCV	2	SCAD
Not inserted in Müller *et al.* (2013)							

TABLE 7.1: Settings of LMM selection procedures with shrinkage. "Reference" refers to the initials of the authors followed by the second digit of the year of publication. The second, the third and the forth columns contain the information about the desired properties for the - fixed and/or random - estimators: consistency and the "oracle properties" (sparsity and asymptotic normality, Fan and Li (2001). The symbol * is added to the authors that proved the oracle properties only for the fixed effects.

It is worth noting that, all simulations are applied with a moderate number of random effects (for both the full and the true model) and of variance-covariance parameters, except for that of Li *et al.* (2018) and Ahn *et al.* (2012). A large amount of fixed effects occur in the full model of Chen *et al.* (2015), Ghosh and Thoresen (2018) and Rohart *et al.* (2014).

To determine the set of candidate models for $\boldsymbol{\beta}$, $|M_{\boldsymbol{\beta}}|$, the authors do not follow the same criterion. Some authors focus only on covariates and in this sense $|M_{\boldsymbol{\beta}}|$ is equal to 2^{p-1} (so the intercept is not included for size of $\boldsymbol{\beta}$). Others instead refer to p as the whole fixed regression parameters, including the intercept, and thus the candidate models are 2^p. Furthermore some authors, such as Kawakubo *et al.* (2014), state that they exclude from $|M_{\boldsymbol{\beta}}|$ the null model (i.e. the model containing only the intercept).

Kawakubo and Kubokawa (2014) found that both the McAIC and a model averaging procedure (which has more appropriate weights) depending on McAIC, work better than cAIC in terms of prediction errors. They prove empirically the same results in the case of small area prediction, which is the topic on which Kawakubo *et al.* (2014) and Lombardía *et al.* (2017) focus on. They show, therefore, a prediction error improvement of CScAIC with respect to cAIC. Compared to mAIC, cAIC and BIC, the EBIC of Kubokawa and Srivastava (2010) is the criterion which, by simulation, leads to a better selection of the true model as the number of covariates and the number of clusters increase. These results constitute empirical evidence of the consistency property of the EBIC. Lombardía *et al.* (2017), instead, compared the extended generalized AIC they defined (7.20) with the conditional AIC defined by Vaida and Blanchard (2005). They discovered that the xGAIC for the Fay-Herriot model presents better performances in terms of correct classification rates of the true model. As the number of covariates increases, the xGAIC performs better and better (in a scenario with three variables it perfectly brings to the correct model), instead the vAIC selects 44% of the times a model with a fewer number of fixed effects. Wenren and Shang (2016) show that the proposed conditional criteria perform more efficiently than the classic Mallow's C_p when more significant fixed effects are added. A large number of units for each cluster is required, if one works with the random effects within clusters (for instance small area estimation) or if one could obtain a less biased estimation of the penalty term. Wenren *et al.* (2016) show by simulation that their two marginal C_p-types perform better, in selecting the correct model, than mAIC and mBIC in particular situations: when observations are few and highly correlated or when the true model

is included in all candidate models and includes more significant fixed effect variables.Kuran and Özkale (2019) compare the performance of their conditional ridge C_p with the CC_P of Wenren and Shang (2016), in both cases of known and unknown variance-covariance matrices of the random effects and of the random errors. Furthermore, they use different values for the ridge parameters and compare various models (with different number of the explicative variables). They show that the percentages of choosing the true model by all the C_p statistics are quite optimal and comparable and they increase as the number of fixed effects increases as well. When the ridge parameter increases, the number of individuals and the number of units are quite small and the correlation between explanatory variables is not high, the CRC_p outperformes the CC_p.

Focusing on the shrinkage selection procedures, Hossain *et al.* (2018) compare the performances, in terms of mean squared prediction errors, reached by their PT and PSE estimators against the unrestricted MLE, the restricted MLE, the LASSO and ALASSO methods. They show that their methodology, as the sample size increases and the number of active covariates decreases, brings to better performance than the other estimators except the restricted MLE. Ghosh and Thoresen (2018) try to demonstrate the great performances of the SCAD penalty over ℓ_1 penalization. Hence, by simulations, they point out that both in a low dimensional setting and in a high dimensional setting the two penalties correctly select the true fixed effects. With respect to ℓ_1, SCAD focuses on a smaller activate set of $\boldsymbol{\beta}$, especially, in the high-dimensional case. Marino *et al.* (2017) compare their penalized likelihood procedure for multilevel models with missing models with the LASSO method applied on data without missing values and, hence, used as benchmark reference. Therefore they also compare the performance of their method with the regularized LASSO on complete-case data. When missing data are present in the dataset the proposed methodology performs better, especially when the number of imputations increases. Taking into account only one imputation doesn't produce huge benefits. On the other hand, the methodology is quite good in identifying the correct model when the number of imputation and the number of units increases. Rohart *et al.* (2014) reached the same results as Schelldorfer *et al.* (2011) in the case of known variances, but with an algorithm much faster. It is worth noting that their method can be computationally combined with other procedures. The orthogonal-based SCAD procedure of Wu *et al.* (2016) is very efficient in selecting the fixed effects as the number of total units increases, but has to be improved for the selection of the random effects. Pan (2016) compared the ability of his two-stage procedure to correctly identify the two kinds of effects

with that of Ahn *et al.* (2012) and Bondell *et al.* (2010). He found that the percentage of the effects (taken both separately and together) correctly identified was higher than the others and was rose as the number of clusters increased. Only in the case of a non normal distribution assumed for ϵ did the method proposed by Ahn *et al.* (2012) perform better, since it does not need any distributional assumptions. Pan (2016) also compares the computational efficiency of his model selection with that of Bondell *et al.* (2010), and concludes that his algorithm takes less time to converge. There are two probable reasons: σ^2 is not included in the profile log- likelihood used by Pan (2016) and a two stage procedure for selecting both the effects is faster than the procedures involving only one step. Lin *et al.* (2013) used the same settings for their simulations as those used by Bondell *et al.* (2010), that is the reason why their results are missing in Table 7.2: they are available in Table 2 of Müller *et al.* (2013). The robust selection method presented by Fan *et al.* (2014) has been shown to lead to the same results of the equivalent non robust method if the data do not present outliers. On the other hand, the method has no influence on the estimates if outliers are present in the data (both in the response variable and in the covariates), while the non robust methodology brings to overfitting with lower fit percentages and higher mean squared errors of the estimated parameters as a consequence. The robust selection method is perturbed by outliers if these are only in the response variable or in the covariates.

In the case of high-dimensional settings where the focus is on selection the fixed and the random effects, Li *et al.* (2018) used in their simulations two ways of controlling the tuning parameters: a non-adaptive regularization (NAR), which chooses the tuning parameter from a simple grid of values, and an adaptive regularization (AR), which attributes weights to different penalty parameters. The AR methodology leads to smaller estimation bias for the variance components and to a better control of the false discovery rate. Chen *et al.* (2015) obtained a good performance selection in terms of low proportion of parameters that didn't shrink to zero while one expected the opposite or of parameters shrinking to zero, by mistake. Furthermore, they obtained accurate results in terms of bias and standard deviations of the estimates. They conducted some simulations excluding from the selection the fixed effects and they discovered that in all situations the fixed effect selection never affects the power performances.

The parameter subset selection method proposed by Schmidt and Smith (2016) leads to better performances, compared to other techniques, among which LASSO, ALASSO and M-ALASSO.

As specified at the beginning of this review, our purpose is to give a clear outline of most methodologies used in linear mixed models that are available in the literature. Hence, in this sense, Table 7.3 summarises all the features that easily identify all procedures: the part of the model focusing on (fixed and/or random), the dimension of the linear mixed model used and the structure of variance and covariance matrices. Dimensionality represents the level of the number of parameters ($\boldsymbol{\theta} = \boldsymbol{\beta}, \boldsymbol{\tau}$) involved in the model. We included not only the methods mentioned by this article, but also those contained in Müller *et al.* (2013), in order to provide a global view of all methodologies. Taking a look jointly to Table 2 of Müller *et al.* (2013), Table 7.2 and Table 7.3, it becomes obvious that most model selection procedures, focusing on selecting both the fixed and the random part in cases of medium and/or high dimensionality, involve a shrinkage procedure. The shrinkage methods are computationally more efficient and statistically accurate (Bühlmann and van de Geer, 2011; Müller *et al.*, 2013).

7.8 Review of real examples

LMM are widely used in medical statistics and biostatistics. To enrich this review, we give a brief look at the real examples described in some of the listed papers.

Ahn *et al.* (2012), Pan (2016) and Hossain *et al.* (2018) describe the Amsterdam Growth and Health Study, widely used in literature. The Amsterdam Growth and Health Study Data were collected to explore the relationship between lifestyle and health in adolescence and young adulthood. In growing towards independence, the lifestyle habits of teenagers change substantially with respect to physical activity, food intake, tobacco smoking, etc. Accordingly, their health perspective may also change. Individual changes in growth and development can be studied by observing and measuring the same participant over a long period of time. The Amsterdam growth and health longitudinal study was designed to monitor the growth and health of teenagers and to develop future effective interventions for adolescence. A total of 147 subjects in the Netherlands participated in the study, and they were measured over 6 time points, thus the total number of observations is 882. The continuous response variable of interest was the total serum cholesterol expressed in mmol/l. Pan (2016) in his paper analyses a second dataset, which is the Colon Cancer Data. The goal of the analysis was to estimate the cost attributable to colon cancer after initial diagnosis by cancer stage,

comorbidity, treatment regimen, and other patient characteristics. The data reported aggregate Medicare spending on a cohort of 10,109 colon cancer patients up to 5 years after initial hospitalization, and these data are considered as the response for a linear mixed model.

Taylor *et al.* (2012) applied their method to determine quantitative trait loci (QTL) in a wheat quality data set. The data set was obtained from a two-phase experiment conducted in 2006 involving a wheat population consisting of 180 double haploid (DH) lines from the crossing of two favoured varieties. Data were collected from two phases of experimentation consisting of an initial field trial and milling laboratory experiment. A partially replicated design approach was used at both experimental phases. The field trial was designed as a randomised block design. The analysis considers a very large set of candidate variables, and matrix $\mathbf{a}$ in Equation (7.59) is a (390×1) size matrix.

Jiang *et al.* (2008) considered a dataset from a survey conducted in Guatemala regarding the use of modern prenatal care for pregnancies where some form of care was used. They consider applying the fence method in selection of the fixed covariates in the variance component logistic model. Again, they cope with a quite large number of covariates.

Marino *et al.* (2017) worked on a dataset provided by the Healthy Directions–Small Business study conducted by Sorensen *et al.* (2005). Some recent epidemiological studies proved that there is a relationship between dietary patterns and physical inactivity with multiple cancers and chronic diseases. One of the main purposes of the study was to detect whether or not the cancer prevention (based on occupational health and health promotion) could lead to reduce significantly the red meat consumption or to improve significantly the mean consumption of fruits and vegetables, the levels of physical activity, the smoking cessation and the reduction of occupational carcinogens. The HD-SB study was a randomized, controlled trial study conducted between 1999 and 2003 as part of the Harvard Center Prevention Program Project. The study population of the study were twenty-six small manufacturing worksites that employed multi-ethnic, low-wage workers. Participating worksites were randomized to either the 18-month intervention group or minimal intervention control group. The respondents to the study were 974 but only 793 of them answered with complete information, hence there was 18.5% of missing data. The number of variables involved in the survey was huge and they were grouped according different areas: health behaviors, red meat consumption, physical activity and consumption of multivitamin and sociodemographic characteristics. The authors took into account 15 covariates and they built a linear-mixed model where the mean consumption of fruit and

vegetables at follow-up. They proposed their methodology for missing data with $1, 3$, and 5 imputations, comparing the results to the analysis made on the complete-cases data.

Fan *et al.* (2014) applied their robust method on a longitudinal progesterone dataset, available on Diggle P.J.'s homepage: https://www.lancs. The dataset contains 492 urine samples from 34 women in a menstrual, where each woman contributed from 11-28 times. The menstrual cycle length was standardized for all women to a reference 28-day cycle. A linear mixed-model was analysed by the authors with the log-transformed progesterone level as response variable, a random intercept and 7 fixed effects: age, bmi, time, the squared values of time and the three first-level interactions among age, bmi and time.

Li *et al.* (2018) in their paper analyze two datasets. The first is related to a longitudinal randomized controlled trial, involving 423 adolescent children from an Hispanic population in New York City had their parents affected by HIV+. The main purpose was to investigate about a negative state of mind (measured by a Basic Symptoms Inventory, a score well described by Weiss (2005)), over six years (each person has been visited about 11.5 times). Six variables were involved in the original dataset, i.e.: treatment (or control group), age, gender, Hispanic (1=Yes, 0=No), visit time (expressed in logarithm of year) and visit season. The authors, worked on a linear mixed model containing the six covariates plus the two-way interactions between treatment and time, gender and Hispanic, counting so 10 predictors, which were included in all the two types of effects. Their regularization procedure was applied both with the non-adaptive version and with the adaptive version (through the inverse of the estimated from the ridge-penalization procedure). Their second dataset is related to a clinical study that investigated on a possible relationship of some protein signatures with post-transplant renal functions for people with a kidney transplant. The study involved 95 renal transplant patients. The main purpose of the study was to analyze which proteins had a significant influence on the longitudinal trajectory of renal function measured by glomerular filtration rate (GFR) of the patients.

Lombardía *et al.* (2017) analyzed a dataset about surveys conducted from the behavioural risk factors information system in Galicia (2010-2011). The sample design applied in the survey was a stratified random sampling, allocating with equal proportions by sex and age group. Forty-one areas from the 53 counties in Galicia were involved in the survey. The authors tried to estimate the prevalence of smokers (at least 16 years old) distinguished by sex. The minimum sample size in the domain was 44 for men and 48 for women. The response variable, employed in the

Fay-Herriot model used, was the logarithmic transformation of smokers'numbers. The covariates were globally 14, classified in four groups: age, degree of urbanization, activity and educational level.

Han (2013) analyzed a public health dataset about obesity released by the U.S. Centers for Disease Control and Prevention, which realized a large health study (6971 people) in the United States (51 counties of California) in the years between 2006 and 2010. The information obtained by the surveys. The purpose of the author was to estimate county level obesity rates for the female Hispanic population within working ages of 18-64.

Bondell *et al.* (2010) consider a recent study of the association between total nitrate concentration in the atmosphere (TNO3,ug/m^3) and a set of measured predictors. Nitrate is one of the major components of fine particulate matter (PM2.5) across the United States. However, it is one of the most difficult components to simulate accurately using numerical air quality models. Identifying the empirical relationships that exist between nitrate concentrations and a set of observed variables that can act as surrogates for the different nitrate formation and loss pathways can help the research and can allow for more accurate simulation of air quality. To formulate these relationships, data obtained from the U.S. EPA Clean Air Status and Trends Network (CASTNet) sites are used. The CASTNet dataset consists of multiple sites with repeated measurements of pollution and meteorological variables on each site, i.e.: the mean ambient particulate ammonium concentration (NH4,ug/m^3), the mean ambient particulate sulfate concentration (SO4,ug/m^3), relative humidity (RH,%), ozone (O3,ppb), precipitation (P,mm/h), solar radiation (SR,W/m^2), temperature (T, °C), temperature difference between 9 m and 2 m probes (TD,°C) and scalar wind speed (WS, m/s). The same data were used by Li *et al.* (2014) to apply their proposed MDL procedure. A subset of the CASTnet dataset was, instead, implied by Chen *et al.* (2015), who focused only on five sites across the eastern United States, (2001-2009) and took as original variables TNO3, NH4 and SO4, instead the others variables were transformed from ours to seasonal, substituting the maximum value for O3 and the mean value for the others. The total number of observations were 175 and in the two-way random effect model the variable time and sites were included as main random effect.

Ghosh and Thoresen (2018) investigated the effects of intake of oxidized and non-oxidized fish oil on inflammatory markers in a randomized study of 52 subjects (dataset already studied in literature). Inflammatory markers were measured at baseline and after three and seven

weeks. They use the data to investigate whether there are any associations between gene expressions measured at baseline and level of the inflammatory marker ICAM-1 throughout the study. From a vast set of genes, they initially selected $p = 506$ genes having absolute correlation greater than or equal to 0.2 with the response at any time point, so that the total number of fixed effects considered becomes $p = 512$. On the other hand, removing the missing observations in the response variable they obtain $n = 150$ observations, making it a high-dimensional selection problem. Further, due to the longitudinal structure of the data, they additionally considered random effect components in the model: they included random intercept and a random slope.

Finally, Rohart *et al.* (2014) apply their approach to a real data set from a project in which hundreds of pigs were studied, the aim being to shed light on the relationships between some of the phenotypes of interest and metabolic data. Linear mixed models are appropriate in this case because observations are in fact repeated data collected in different environments (groups of animals reared together in the same conditions). Some individuals were also genetically related, introducing a family effect. The data set consisted of 506 individuals from 3 breeds, 8 environments and 157 families, metabolic data contained p = 375 variables, and the phenotype investigated was the Daily Feed Intake (DFI).

Li and Zhu (2013) applied their new covariance-based test on a famous pig weight dataset, containing the weights of 48 pigs, measured in 9 successive weeks.

Reference	Model	m/n_i	p/p_f	s_i/s_{f_i}	q/q_f	$\lvert M_\beta\rvert/\lvert M_\tau\rvert$	$min\lvert\beta_k\rvert/\sigma$	$min\{ev(\Psi/\sigma^2)\}$	e/u	Method
AZL12 (Ahn *et al.*, 2012)	int + slope	(100, 200)/(5,10)	3/9	(2,4)/(5,10)	(4,7)/(16,56)	511/1024	(0.8,0.62)	(0.5,0.3)	N/(N,t_5,Exp)	shrinkage
CLSZ15 (Chen *et al.*, 2015)	int	30/5,40/6,60/8	3/(30,60,100)	1/1	3/3	$2^{30/60/100}/1$	2		$\sqrt{0.75}t_8/N$	shrinkage
FQZ14 (Fan *et al.*, 2014)	int+slope	50/5	3/8	3/4	7/11	256/	1.5	0.5	N/N	shrinkage
GT16 (Ghosh and Thoresen, 2018)	int+slope	25/6	5/(10,50,300,500)	2/2	2/2	/1	2	0.56	N	shrinkage
HTA18 (Hossain *et al.*, 2018)	int+slope	(40,75,150)/5	5/(7,10,14,19)	2/2	2/2	/1	1.84	1.26	N/N	shrinkage
KK14 (Kawakubo *et al.*, 2014)	int	10/50	5/7	1/1	1/1	7/1	0.35	1	N	McAIC
KO18 (Kuran and Özkale, 2019)	int	(10,20,30)/(10,20)	(2,3,4)/5	1/1	2/2	15/1	2		N/N	ridge CC_p
KS10 (Kubokawa and Srivastava, 2010)	int	(6,15)/4	(2,4,6)/7	1/1	2/2	7/1	2	0.1	N/N	EBIC
KSK14 (Kawakubo *et al.*, 2014)	int	30/3	3/5	1/1	2/2	31/1	1	1	N	CScAIC
LLVR17 (Lombardía *et al.*, 2017)	int	53/41	(1,2,3)/14	1/1	2/2	3/1			N/N	xGAIC
LPJ13 (Lin *et al.*, 2013)	int+slope	(30,60/5,10)	2/9	3/4	7/11	512/16	1	0.45	N	shrinkage
LWSWZZ18 (Li *et al.*, 2018)	int+slope	200/8,100/5	5/101-601	4/51	$11/\frac{51*50}{2}$	$2^{600/100}/2^{50}$	1		N/N	shrinkage
LYCZ14 (Li *et al.*, 2014)	int+slope	(50,80)/4	(4,7,10,13)/13	3/3	7/7	13/1	0.2	0.10	N	MDL
LZ13 (Li and Zhu, 2013)	int+slope	(40,70,100)/Poisson(3)+2	2/2	2/2	3/3				$t_6/\sqrt{(15)}$/*Bivariate N*	difference-based test
MBL17 (Marino *et al.*, 2017)	int	40/5,(60,150)/25	3/8	1/1	2/2	256/1	1.5		N/N	shrinkage
P16 (Pan, 2016)	int + slope	(50,100,200)/5	3/5	2/5	4/15	255/31	(1.5,1.06)	0.34	N/N	shrinkage
P16 (Pan, 2016)	int + slope	(30,60)/(5,10)	2/9	3/4	7/11	511/15	1	0.45	N/N	shrinkage
P16 (Pan, 2016)	int + slope	(30,60,90)/12	3/8	3/5	7/16	255/31	(1.5,1.06)	0.34	N/N	shrinkage
PS18 (Pan and Shang, 2018b)	int+slope	(30/5),(60/10),(10,20/10)	2/9,3/5	3/4,2/4	7/11,4/11					shrinkage
RSCL14 (Rohart *et al.*, 2014)	int+slope	(20,15)/(6,8)	5/(80,300,600)	(2,3)/(2,3)	4/11	8/16	1	0.43	N	shrinkage
SS16 (Schmidt and Smith, 2016)	int+slope	30/5	2/9	3/4	7/11	512/2	1	0.44	N/N	subset selection
TVCN12 (Taylor *et al.*, 2012)	int+slope	(10, 20)/10	1/(20,40,80)	1/(20,40,80)	(2,3)/(2,3)				N	shrinkage
WLXZ16 (Wu *et al.*, 2016)	int+slope	(30,60)/(5,10)	2/9	3/4	7/11	512/16	1	0.45	N	shrinkage
WLXZ16 (Wu *et al.*, 2016)	int+slope	(10,20)/10	3/5	2/4	4/11	16/16	0.5	0.43	N,t3,t copula	shrinkage
WLXZ16 (Wu *et al.*, 2016)	int+slope	40/10	2/4	2/4	4/11	16/8	1	0.43	N	shrinkage
WS16 (Wenren and Shang, 2016)	int	(5,20,50)/5	(2,3,4)/5	1/1	2/2, 0/0	15/1	2	1	N	CC_p
WSP16 (Wenren *et al.*, 2016)	int	5/(5,10,20)/5	(2,3,4)/5	1/1	1/1	16/1	2	1	N	MC_p

TABLE 7.2: Table containing the information of simulations, following the same setting of Müller et al. (2013) where "Reference" refers to the initials of the surnames of the authors followed by the second digits of the year of publication; m and n_i are the number of clusters and the number of units per cluster; p and p_f the number of the fixed parameters in the true model and in the full one

7.9 Discussion and conclusion

In this paper we have discussed most of the model selection procedures for linear mixed models available to date. The purpose of our review is to allow users to easily identify the type of method they need, according to certain characteristics, such as the number of clusters and/or the number of units per cluster, the part of the model to be selected (fixed and/or random), the dimension of the model and the structure of the variance-covariance matrices. For all the methods, a description of the simulations, if available, is reported in Table 7.2: the purpose is to give an idea of the model settings and not to provide evidence of the best methods. We used more or less the same notation as Müller et al. (2013) for alignment with the previous review and, hence, facilitating the comparisons of the various methods over time. But this review is not only an update of Müller's review (Müller *et al.*, 2013), but an attempt to cluster the procedures from a different point of view: the part of the model to be selected, fixed and/or random. As a matter of fact, this is one of the main issues when looking for an appropriate method to choose. Moreover, particular attention is given to the SW used, together with the implementation and the availability of the code.

This review mentions the available theoretical properties corresponding to the different methodologies, with comparisons among them whereas it's possible. A relevant importance is given here to the shrinkage methods (focused on the selection of fixed and/or random effects), since these procedures need for the oracle properties established by Fan and Li (2001).

By simulation the authors considered in this review try to achieve the best result, i.e. to identify the optimal model among a pool of candidate models and not the true model. Many issues are related to the choice of the optimal model, one of which is determined by the dimension of the pool of candidate models (2^{p+s}). The larger this set $\mathcal{M}$, the lower computational efficiency is. This has been proven by Fence methods and a number of Bayesian methods reported in (Müller *et al.*, 2013) as well as the two stage procedures of Section 7.6.2, which select the two effects separately, thus reducing the overall dimension of models.

Over time, greater attention has been given to the generalization of $\boldsymbol{\Sigma}$ in Equation (7.2): the scaled version $\sigma^2\boldsymbol{\Sigma}_*$ replaced $\sigma^2 I_{n_i}$, but except for Shang and Cavanaugh (2008) the scaled version $\sigma^2\boldsymbol{\Psi}_*$ is assumed for $\boldsymbol{\Psi}$. There is still poor theoretical support for a generalized scenario of the variance-covariance matrices for both the effects.

Most of the methods were implemented in R, using different packages or through their own codes (not published in any package). Some

Reference	Focus	Dimensionality	Ψ	Σ	Software
BKG10 (Bondell *et al.*, 2010)	Fixed+Random	Low/Medium	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	R
CD03 (Chen and Dunson, 2003)	Random	Low	Ψ	$\sigma^2 I_{n_i}$	
DMT11 (Dimova *et al.*, 2011)	Fixed+Random	Low	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	
GK10 (Greven and Kneib, 2010)	Random	Low	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	"cAIC4"R package
IZGG11 (Ibrahim *et al.*, 2011)	Fixed+Random	Low/Medium	Ψ	$\sigma^2 I_{n_i}$	R
JNR09 (Jiang *et al.*, 2009)	Fixed	Low	σ^2	σ^2	"fence"R package
JR03 (Jiang and Rao, 2003)	Fixed+Random	Low	Ψ	Σ	
JRGN08 (Jiang *et al.*, 2008)	Fixed	Medium	Ψ	Σ	"fence"R package
K11 (Kubokawa, 2011)	Fixed + Random	Low	Ψ	Σ	
NJ12 (Nguyen and Jiang, 2012)	Fixed	High	$\sigma_b^2 I$	$\sigma_\epsilon^2 I$	"fence"R package
PL12 (Peng and Lu, 2012)	Fixed+Random	Low/Medium	Ψ	$\sigma^2 I_{n_i}$	Matlab
PN06 (Pu and Niu, 2006)	Fixed+Random	Low	Ψ	Σ	
SC08 (Shang and Cavanaugh, 2008)	Fixed+Random	Low	Ψ	$\sigma^2\Sigma_*$	
SK10 (Srivastava and Kubokawa, 2010)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	
Inserted in (Müller *et al.*, 2013))					
Reference	**Focus**	**Dimensionality**	Ψ	Σ	
AZL12 (Ahn *et al.*, 2012)	Fixed+Random	Low/Medium	Ψ	$\sigma^2 I_{n_i}$	
CLSZ15 (Chen *et al.*, 2015)	Fixed+Random	High	$\sigma_i^2 I_{n_i}$	$\sigma^2 I_{n_i}$	
FQZ14 (Fan *et al.*, 2014)	Fixed+Random	Low	$\sigma^2 I_{ni}$	$\sigma^2 I_{ni}$	
GT16 (Ghosh and Thoresen, 2018)	Fixed	Low/High	Ψ	$\sigma^2 I_{n_i}$	R
H13 (Han, 2013)	Fixed	Low/Medium	$\sigma_b^2 I_{n_i}$	$\sigma_i^2 I_{n_i}$	R
HTA18 (Hossain *et al.*, 2018)	Fixed	Low/Medium	Ψ	Σ	
KK14 (Kawakubo and Kubokawa, 2014)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2\Sigma_*$	
KO18 (Kuran and Özkale, 2019)	Fixed	Low/Medium	$\sigma^2\Psi_*$	$\sigma^2 I_{ni}$	R
KS10 (Kubokawa and Srivastava, 2010)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2\Sigma_*$	
KSK14 (Kawakubo *et al.*, 2014)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2\Sigma_*$	
LLVR17 (Lombardía *et al.*, 2017)	Fixed	Low/Medium	Ψ	Σ	R
LPJ13 (Lin *et al.*, 2013)	Fixed+Random	Medium	Ψ	$\sigma^2 I_{n_i}$	R
LS15 (Lahiri and Suntornchost, 2015)	Fixed	Low/Medium	$\sigma_{b_i}^2$	$\sigma^2 I_{n_i}$	
LWSWZZ18 (Li *et al.*, 2018)	Fixed+Random	High	$\sigma^2\Psi_*$	$\sigma^2 I_{ni}$	
LYCZ14 (Li *et al.*, 2014)	Fixed	Low	Ψ	$\sigma^2 I_{n_i}$	
LZ13 (Li and Zhu, 2013)	Random	Low/(Medium)	Ψ	$\sigma^2 I_{n_i}$	
MBL17 (Marino *et al.*, 2017)	Fixed	Low(Medium)	$\sigma^2\Psi_*$	$\sigma_i^2 I_{n_i}$	
P16 (Pan, 2016)	Fixed+Random	Low/Medium/High	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	R
PS18 (Pan and Shang, 2018b)	Fixed+Random	Low/Medium	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	R
RSCL14 (Rohart *et al.*, 2014)	Fixed(+Random)	High	Ψ	$\sigma^2 I_{n_i}$	"MMS"R package
SS16 (Schmidt and Smith, 2016)	Fixed+Random	Low/(Medium)	Ψ	$\sigma^2 I_{n_i}$	Matlab
TVCN12 (Taylor *et al.*, 2012)	Fixed+Random	Medium/High	$\sigma^2\Psi_*$	$\sigma^2\Sigma_*$	ASReml-R
WLXZ16 (Wu *et al.*, 2016)	Fixed+Random	Low/Medium	Ψ	$\sigma^2 I_{n_i}$	R and Matlab
WS16 (Wenren and Shang, 2016)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	R
WSP16 (Wenren *et al.*, 2016)	Fixed	Low	$\sigma^2\Psi_*$	$\sigma^2 I_{n_i}$	R
Not inserted in (Müller *et al.*, 2013)					

TABLE 7.3: Settings of LMM selection procedures for all the procedures analyzed in the review. "Reference" refers to the initials of the authors followed by the second digit of the year of publication (we use the same approach as (Müller *et al.*, 2013)); "Focus" indicates the part of the model that is subject to selection (Fixed, Random or both); "Dimensionality" is inherent to the number of parameters involved in the initial model; $\mathbf{\Psi}$ and $\mathbf{\Sigma}$ describe the structure assumed for the variance-covariance matrices related to the random effects and the random component, respectively; "Software" specifies the software (when specified) used for implementation of the procedure

authors, however, do not even specify the software used (see Table 7.3). As in a Meta-analysis, we gathered the simulations presented in the papers described but, since the results are not directly comparable, the tables synthesise the main parameters characterizing the simulations.

Hence, the main purpose of this review was to provide an overview of some useful components/factors characterizing each selection criterion, so that users can identify which method to apply in a specific situation also. In addition, an effort was made to tidy up the notation used in the literature, by "translating", if necessary, symbols and formulas in each paper into a common "language".

Chapter 8

Conclusions

In this Thesis, at first, we proposed a new position weighted rank correlation coefficient for linear orders. We demonstrated that the proposed coefficient is in one to one correspondence with the weighted Kemeny distance proposed by García-Lapresta and Pérez-Román (2010), when equal importance is assigned to items'positions, the weighted rank correlation coefficient is equivalent to the rank correlation coefficient defined by Emond and Mason (2002).

Then, we provided a weighted rank correlation coefficient τ_x^w for weak orderings, as an extension of τ_x^w for linear orderings. We demonstrated the correspondence between τ_x^w and the weighted Kemeny distance and, finally, we showed that, in the case of tied rankings and $w_i = \frac{1}{m-1}$ for all i, the weighted rank correlation coefficient τ_x^w is equal to the Emond and Mason rank correlation coefficient τ_x. By means of simulations, we demonstrated that a modified BB algorithm allows us to find the true consensus and to verify the effect of the weighting vector. The analysis of two real datasets shows, as demonstrated analytically, that with $w_i = \frac{1}{m-1}$ for all i we obtain the same solution without weightings, while the solutions always differ as soon as we simplify the weighting structure.

Some crucial considerations could represent the basis for future developments: firstly to take into account the multiple solutions of the consensus process, since only one random solution has been considered in this thesis (in order to facilitate the implementation process); then, to focus on the optimization of the implemented procedures in order to achieve faster algorithms; in the end, the development of the same analysis for items'importance for a complete consensus process.

Moreover, we focused on distance-based decision trees for ranking data, when the position occupied by items is relevant. We proposed the weighted Kemeny distance as impurity function and the relative proper weighted correlation coefficient in order to achieve the consensus measure in the terminal nodes. Our methodology found to be capable of

identifying correctly homogeneous groups of rankings when more than one position is taken into account. The implementation of a faster algorithm for the rpart package (Therneau *et al.*, 2010) could lead to work faster in the presence of a large number of items. Further developments could be, hence, a replication of the same analyses with an increasing number of items.

We proposed a combination of multiple decision trees in order to construct more powerful prediction models: Boosting and Bagging (with replacement and with OOB, without replacement), for ranking data. Once the trees are built up, τ_x was employed for assigning the median ranking as the final prediction, tree by tree, and for measuring the relative error. We applied the above methodologies to simulated data and to a real case showing that boosting outperforms bagging (both with and without replacement). By means of simulations the sensitivity of the procedures to the number of trees, the heterogeneity of the data and the depth of the single tree has been studied. The Boosting for ranking data was also extended to the case of different positions'weights structures. Once the trees are built up, τ_x^w was employed for assigning the median ranking as the final prediction, tree by tree, and for measuring the relative error. By means of simulations the sensitivity of the procedures to the different weighted structures was studied, in terms of error and variable importance. The extension to the definition and analysis of Random Forests for ranking data with position weights (and in future with item weights) will be necessary, for the sake of completeness in the framework of Ensemble Methods for ranking data.

The last topic (selection of effects in Linear Mixed Models) could seem far away from the consensus ranking problem, but, actually, we could consider the output of a model selection process as a ranking; therefore, using different measures (AIC, BIC,. . .) which provide different rankings of the models, a consensus ranking process could be applied in order to identify the "optimum"ranking of the models. In a few words, each measure could provide a ranking of linear mixed models and a consensus ranking process could be useful for detecting the best consensus of the ranked models.

Bibliography

Ahn, M., Zhang, H. H., and Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica*, **22**(4), 1539.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. In *Breakthroughs in statistics*, pages 610–624. Springer.

Amodio, S., D'Ambrosio, A., and Siciliano, R. (2016). Accurate algorithms for identifying the median ranking when dealing with weak and partial rankings under the kemeny axiomatic approach. *European Journal of Operational Research*, **249**(2), 667–676.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, **66**(4), 1069–1077.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, **52**, 345–370.

Braun, J., Held, L., and Ledergerber, B. (2012). Predictive cross-validation for the choice of linear mixed-effects models with application to data from the swiss hiv cohort study. *Biometrics*, **68**(1), 53–61.

Breiman, L. (1996). Bagging predictors. *Machine learning*, **24**(2), 123–140.

Breiman, L. *et al.* (1998). Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, **26**(3), 801–849.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). Classification and regression trees. wadsworth int. *Group*, **37**(15), 237–251.

Buscemi, S. and Plaia, A. (2019). Model selection in linear mixed-effect models. *AStA Advances in Statistical Analysis*, pages 1–47.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer.

Chen, F., Li, Z., Shi, L., and Zhu, L. (2015). Inference for mixed models of anova type with high-dimensional data. *Journal of Multivariate Analysis*, **133**, 382–401.

Chen, J., Li, Y., and Feng, L. (2012). A new weighted spearman's footrule as a measure of distance between rankings. *arXiv preprint arXiv:1207.2541*.

Chen, J., Li, Y., and Feng, L. (2014). On the equivalence of weighted metrics for distance measures between two rankings. *Journal of Information & Computational Science*, **11**(13), 4477–4485.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics*, **59**(4), 762–769.

Cheng, W., Hühn, J., and Hüllermeier, E. (2009). Decision tree and instance-based learning for label ranking. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 161–168. ACM.

Cook, W. D. (2006). Distance-based and ad hoc consensus models in ordinal preference ranking. *European Journal of operational research*, **172**(2), 369–385.

Cook, W. D., Kress, M., and Seiford, L. M. (1986). An axiomatic approach to distance on partial orderings. *RAIRO-Operations Research*, **20**(2), 115–122.

D'ambrosio (a), A., Amodio, S., and Iorio, C. (2015). Two algorithms for finding optimal solutions of the kemeny rank aggregation problem for full rankings. *Electronic Journal of Applied Statistical Analysis*, **8**(2), 198–213.

Dimova, R. B., Markatou, M., and Talal, A. H. (2011). Information methods for model selection in linear mixed effects models with application to hcv data. *Computational Statistics & Data Analysis*, **55**(9), 2677–2697.

D'Ambrosio (b), A., Amodio, S., and Mazzeo, G. (2015). Consrank, compute the median ranking (s) according to the kemeny's axiomatic approach. r package version 1.0. 2.

D'Ambrosio, A. and Heiser, W. J. (2016). A recursive partitioning method for the prediction of preference rankings based upon kemeny distances. *Psychometrika*, **81**(3), 774–794.

D'Ambrosio, A., Mazzeo, G., Iorio, C., and Siciliano, R. (2017). A differential evolution algorithm for finding the median ranking under the kemeny axiomatic approach. *Computers & Operations Research*, **82**, 126–138.

Emond, E. J. and Mason, D. W. (2000). *A new technique for high level decision support*. Department of National Defence Canada, Operational Research Division (Corporate, Air Maritime).

Emond, E. J. and Mason, D. W. (2002). A new rank correlation coefficient with application to the consensus ranking problem. *Journal of Multi-Criteria Decision Analysis*, **11**(1), 17–28.

Fan, Y. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, **96**(456), 1348–1360.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of statistics*, **40**(4), 2043–2068.

Fan, Y., Qin, G., and Zhu, Z. Y. (2014). Robust variable selection in linear mixed models. *Communications in Statistics-Theory and Methods*, **43**(21), 4566–4581.

Farnoud, F., Touri, B., and Milenkovic, O. (2012). Novel distance measures for vote aggregation. *arXiv preprint arXiv:1203.6371*.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometric regression tools. *Technometrics*, **35**, 109–148.

Freund, Y. and Schapire, R. E. (1998). Discussion: Arcing classifiers. *The Annals of Statistics*, **26**(3), 824–832.

Freund, Y., Schapire, R. E., *et al.* (1996). Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **1**, 302–332.

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397–416.

García-Lapresta, J. L. and Pérez-Román, D. (2010). Consensus measures generated by weighted kemeny distances on weak orders. In *2010 10th International Conference on Intelligent Systems Design and Applications*, pages 463–468. IEEE.

Ghosh, A. and Thoresen, M. (2018). Non-concave penalization in linear mixed-effects models and regularized selection of fixed effects. *AStA Advances in Statistical Analysis*, **102(2)**, 179–210.

Gilmour, S. G. (1996). The interpretation of mallows' cp statistic. *The Statistician*, **45**, 49–56.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, pages 359–378.

Good, I. (1980). C59. the number of orderings of n candidates when ties and omissions are both allowed. *Journal of statistical computation and simulation*, **10**(2), 159–159.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional aic in linear mixed models. *Biometrika*, pages 773–789.

Han, B. (2013). Conditional akaike information criterion in the fay–herriot model. *Statistical Methodology*, **11**, 53–67.

Hansen, M. H. and Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Statistics and Science: A Festschrift forTerry Speed*, pages 145–163.

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, **27**(2), 83–85.

Henzgen, S. and Hüllermeier, E. (2015). Weighted rank correlation: a flexible approach based on fuzzy order relations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 422–437. Springer.

Hodges, J. S. and Sargent, D. J. (2001). Counting degrees of freedom in hierarchical and other richly-parameterised models. *Biometrika*, pages 367–379.

Hossain, S., Thomson, T., and Ahmed, E. (2018). Shrinkage estimation in linear mixed models for longitudinal data. *Metrika*, **81**(5), 569–586.

Hui, F. K., Müller, S., and Welsh, A. (2017). Joint selection in mixed models using regularized pql. *Journal of the American Statistical Association*, **112**(519), 1323–1333.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, **67**(2), 495–503.

Irurozki, E., Calvo, B., Lozano, J. A., *et al.* (2016). Permallows: An r package for mallows and generalized mallows models. *Journal of Statistical Software*, **71**(i12).

Jiang, J. and Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhyā: The Indian Journal of Statistics*, pages 23–42.

Jiang, J., Rao, J. S., Gu, Z., Nguyen, T., *et al.* (2008). Fence methods for mixed model selection. *Annals of Statistics*, **36**(4), 1669–1692.

Jiang, J., Nguyen, T., and Rao, J. S. (2009). A simplified adaptive fence procedure. *Statistics and probability Letters*, **79**, 625–629.

Kawakubo, Y. and Kubokawa, T. (2014). Modified conditional aic in linear mixed models. *Journal of Multivariate Analysis*, **129**, 44–56.

Kawakubo, Y., Sugasawa, S., Kubokawa, T., *et al.* (2014). Conditional aic under covariate shift with application to small area prediction. Technical report, CIRJE, Faculty of Economics, University of Tokyo.

Kawakubo, Y., Sugasawa, S., and Kubokawa, T. (2018). Conditional akaike information under covariate shift with application to small area estimation. *Canadian Journal of Statistics*, **46**(2), 316–335.

Kemeny, J. G. and Snell, L. (1962). Preference ranking: an axiomatic approach. *Mathematical models in the social sciences*, pages 9–23.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, **30**(1/2), 81–93.

Kubokawa, T. (2011). Conditional and unconditional methods for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **102**(3), 641–660.

Kubokawa, T. and Srivastava, M. S. (2010). An empirical bayes information criterion for selecting variables in linear mixed models. *Journal of the Japan Statistical Society*, **40**(1), 111–131.

Kumar, R. and Vassilvitskii, S. (2010). Generalized distances between rankings. In *Proceedings of the 19th international conference on World wide web*, pages 571–580. ACM.

Kuran, Ö. and Özkale, M. R. (2019). Model selection via conditional conceptual predictive statistic under ridge regression in linear mixed models. *Journal of Statistical Computation and Simulation*, **89**(1), 155–187.

Lahiri, P. and Suntornchost, J. (2015). Variable selection for linear mixed models with applications in small area estimation. *Sankhya B*, **77**(2), 312–320.

Lee, P. H. and Yu, P. L. (2010). Distance-based tree models for ranking data. *Computational Statistics & Data Analysis*, **54**(6), 1672–1682.

Li, L., Yao, F., Craiu, R. V., and Zou, J. (2014). Minimum description length principle for linear mixed effects models. *Statistica Sinica*, pages 1161–1178.

Li, Y., Wang, S., Song, P. X.-K., Wang, N., Zhou, L., and Zhu, J. (2018). Doubly regularized estimation and selection in linear mixed-effects models for high-dimensional longitudinal data. *Statistics and its interface*, **11**(4), 721.

Li, Z. and Zhu, L. (2013). A new test for random effects in linear mixed models with longitudinal data. *Journal of Statistical Planning and Inference*, **143**(1), 82–95.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional aic for linear mixed-effects models. *Biometrika.*, pages 773–778.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by reml and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, **22**(2), 341–355.

Liski, E. P. and Liski, A. (2008). Model selection in linear mixed models using mdl criterion with an application to spline smoothing. In *Proceedings of the First Workshop on Information Theoretic Methods in Science and Engineering, Tampere, Finland*, pages 18–20.

Liu, X.-Q. and Hu, P. (2013). General ridge predictors in a mixed linear model. *Statistics*, **47**(2), 363–378.

Lombardía, M. J., López-Vizcaíno, E., and Rueda, C. (2017). Mixed generalized akaike information criterion for small area models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **180**(4), 1229–1252.

Marcus, P., Heiser, W., and D'Ambrosio, A. (2013). *Comparison of heterogeneous probability models for ranking data*. Ph.D. thesis, Master thesis. http://www. math. leidenuniv. nl/scripties/1MasterMarcus. pdf.

Marhuenda, Y., Molina, I., and Morales, D. (2013). Small area estimation with spatio-temporal fay–herriot models. *Computational Statistics and Data Analysis*, **58**, 308–325.

Marino, M., Buxton, O. M., and Li, Y. (2017). Covariate selection for multilevel models with missing data. *Stat*, **6**(1), 31–46.

Marshall, E. C. and Spiegelhalter, D. J. (2003). Approximate cross-validatory predictive checks in disease mapping models. *Statistics in Medicine*, **22**, 1649–1660.

Müller, S., Scealy, J. L., Welsh, A. H., *et al.* (2013). Model selection in linear mixed models. *Statistical Science*, **28**(2), 135–167.

Nguyen, T. and Jiang, J. (2012). Restricted fence method for covariate selection in longitudinal data analysis. *Biostatistics*, **13**(2), 303–314.

Özkale, M. R. and Can, F. (2017). An evaluation of ridge estimator in linear mixed models: an example from kidney failure data. *Journal of Applied Statistics*, **44**(12), 2251–2269.

Pan, J. (2016). *Adaptive LASSO For Mixed Model Selection via Profile Log-Likelihood*. Ph.D. thesis, Bowling Green State University.

Pan, J. and Shang, J. (2018a). Adaptive lasso for linear mixed model selection via profile log-likelihood. *Communications in Statistics-Theory and Methods*, **47**(8), 1882–1900.

Pan, J. and Shang, J. (2018b). A simultaneous variable selection methodology for linear mixed models. *Journal of Statistical Computation and Simulation*, **88**(17), 3323–3337.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, **109**, 109–129.

Philip, L., Wan, W. M., and Lee, P. H. (2010). Decision tree modeling for ranking data. In *Preference learning*, pages 83–106. Springer.

Piccarreta, R. (2010). Binary trees for dissimilarity data. *Computational Statistics & Data Analysis*, **54**(6), 1516–1524.

Plaia, A. and Sciandra, M. (2019). Weighted distance-based trees for ranking data. *Advances in Data Analysis and Classification*, **13**, 427–444.

Plaia, A., Buscemi, S., and Sciandra, M. (2018a). Consensus among preference rankings: a new weighted correlation coefficient for linear and weak orderings. *submitted*, pages 1–28.

Plaia, A., Sciandra, M., and Buscemi, S. (2018b). Consensus measures for preference rankings with ties: an approach based on position weighted kemeny distance. *Advances in Statistical Modelling of Ordinal Data*, page 171.

Plaia, A., Sciandra, M., and Buscemi, S. (2018c). Weighted and unweighted distances based decision tree for ranking data. In *Book of short Papers SIS 2018*, pages 1–7. Società Italiana di Statistica, Pearson.

Plaia, A., Buscemi, S., and Sciandra, M. (2019a). Ensemble methods for preference structures, with relevance to the rankingsí positions. *Book of Abstract ASMDA 2019, Florence*, page 142.

Plaia, A., Buscemi, S., and Sciandra, M. (2019b). A new position weight correlation coefficient for consensus ranking process without ties. *Stat*, **8**(1), e236.

Pu, W. and Niu, X.-F. (2006). Selecting mixed-effects models based on a generalized information criterion. *Journal of Multivariate Analysis*, **97**(3), 733–758.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, **14 (3)**, 1080–1100.

Rocha, F. M. and Singer, J. M. (2018). Selection of terms in random coefficient regression models. *Journal of Applied Statistics*, **45**(2), 225–242.

Rohart, F., San Cristobal, M., and Laurent, B. (2014). Selection of fixed effects in high dimensional linear mixed models using a multicycle ecm algorithm. *Computational Statistics & Data Analysis*, **80**, 209–222.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Schelldorfer, J., Bühlmann, P., and DE GEER, S. V. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, **38**(2), 197–214.

Schmidt, K. and Smith, R. C. (2016). A parameter subset selection algorithm for mixed-effects models. *International Journal for Uncertainty Quantification*, **6**(5).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Sciandra, M. and Plaia, A. (2018). A graphical model selection tool for mixed models. *Communications in Statistics-Simulation and Computation*, **47**(9), 2624–2638.

Shang, J. and Cavanaugh, J. E. (2008). Bootstrap variants of the akaike information criterion for mixed model selection. *Computational Statistics & Data Analysis*, **52**(4), 2004–2021.

Shih, Y.-S. (2001). Selecting the best splits for classification trees with categorical variables. *Statistics & probability letters*, **54**(4), 341–345.

Singer, J. M., Rocha, F. M., and Nobre, J. S. (2017). Graphical tools for detecting departures from linear mixed model assumptions and some remedial measures. *International Statistical Review*, **85**(2), 290–324.

Sorensen, G., Barbeau, E., Stoddard, A. M., Hunt, M. K., Kaphingst, K., and Wallace, L. (2005). Promoting behavior change among working-class, multiethnic workers: results of the healthy directions—small business study. *American Journal of Public Health*, **95**(8), 1389–1395.

Spearman, C. (1987). The proof and measurement of association between two things. *The American journal of psychology*, **100**(3/4), 441–471.

Srivastava, M. S. and Kubokawa, T. (2010). Conditional information criteria for selecting variables in linear mixed models. *Journal of Multivariate Analysis*, **101**(9), 1970–1980.

Sugiura, N. (1978). Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics A*, **7**, 13–26.

Taylor, J. D., Verbyla, A. P., Cavanagh, C., and Newberry, M. (2012). Variable selection in linear mixed models using an extended class of penalties. *Australian & New Zealand Journal of Statistics*, **54**(4), 427–449.

Therneau, T. M., Atkinson, B., and Ripley, M. B. (2010). The rpart package.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Vaida, F. and Blanchard, S. (2005). Conditional akaike information for mixed-effects models. *Biometrika*, **92**(2), 351–370.

Vigna, S. (2015). A weighted correlation index for rankings with ties. In *Proceedings of the 24th international conference on World Wide Web*, pages 1166–1176. International World Wide Web Conferences Steering Committee.

Wang, J. and Schaalje, G. B. (2009). Model selection for linear mixed models using predictive criteria. *Communications in Statistics-Simulation and Computation*, **38**(4), 788–801.

Wang, W. (2016). Identifiability of covariance parameters in linear mixed effects models. *Linear Algebra and its Applications*, **506**, 603–613.

Weiss, R. E. (2005). *Modeling longitudinal data*. Springer Science & Business Media.

Wenren, C. and Shang, J. (2016). Conditional conceptual predictive statistic for mixed model selection. *Journal of Applied Statistics*, **43**(4), 585–603.

Wenren, C., Shang, J., and Pan, J. (2016). Marginal conceptual predictive statistic for mixed model selection. *Open Journal of Statistics*, **6**(02), 239.

Wu, P., Luo, X., Xu, P., and Zhu, L. (2016). New variable selection for linear mixed-effects models. *Annals of the Institute of Statistical Mathematics*, pages 1–20.

Yilmaz, E., Aslam, J. A., and Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 587–594. ACM.

Zhang, X., Liang, H., Liu, A., Ruppert, D., and Zou, G. (2016). Selection strategy for covariance structure of random effects in linear mixed-effects models. *Scandinavian Journal of Statistics*, **43**(1), 275–291.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, **101**(476), 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, B*, **67**(2), 301–320.