UCLouvain

École polytechnique de Louvain

Recovering weights from comparison results in extensions of BTL model

Authors: **Marine Branders, Astrid Vekemans**
Supervisor: **Julien Hendrickx**
Readers: **Jean-Charles Delvenne, Pierre-Antoine Absil**
Academic year 2021–2022
Master [120] in Mathematical Engineering

Abstract

In this master thesis, we generalize the weighted least squares method developed for the BTL model to models extending the latter one. This is done by extending the Gauss-Markov theorem, which finds the Best Linear Unbiased Estimator of parameters from a linear system with noisy observations. This problem of recovering the parameters of a probability model aims to estimate the quality of the items used for example for user services (streaming platform, e-commerce), sport tournament, sociology, and many others. Observations of such scores are often non-existent or poorly calibrated, as opposed to preference observations. Several preference models based on item scores exist, such as the BTL model, based on a probability $w_i/(w_i + w_j)$ that an item i wins over an item j, and various extensions of this model. Methods have been developed to estimate the vector of scores $\boldsymbol{w} = (w_1, \cdots, w_N)$ based on BTL model. This master thesis aims to generalize the method developed in [HOS20] for BTL to recover item scores of BTL-extended models for comparisons depending on those scores and possibly on additional parameters. This thesis focuses on two pairwise comparisons models allowing ties in the observations and on two triple comparisons models without ties.

The extended weighted least squares method is designed with a focus on the weight matrix and has been justified by theoretical results in the context of linear estimators. For this purpose, Gauss-Markov Theorem for BLUE estimators of observable parameters has been generalized for minimum error estimator with possibly unobservable parameters and non-independent observations. Furthermore, we characterize the decrease of the error and we observe that this method outperforms the least squares method and iterative maximum likelihood for the different BTL-extended models.

Acknowledgments

Our very first thank goes to our promoter, Pr. Julien Hendrickx, for his help and involvement in the research of our topic. We are furthermore thankful for his availability and his wise advises during this year. He was indeed always present to ask us the good questions to keep us to improve in our thinking and our work. We would finally want to thank him for his feedback on this report. We would also like to thank his research team for their help in our work.

Besides, we thank our two readers, Pr. Pierre-Antoine Absil and Pr. Jean-Charles Delvenne, for accepting to read our master thesis and to be a member of our jury.

We also thank our friend, Bastien, for his feedback on this report. In a more personal point of view, we thank our friends and family, present and supportive during those five years of studies. Especially Adrien and Manon, met during our first year at university, and Gauthier, who encouraged us during this thesis.

Contents

Notation

N	Total number of items 2
$\tilde{N}$	Total number of parameters (number of items + number of additional parameters) 33
n	Number of items composing a ranking 5
n_s	Number of sets of n items compared to each other 26
$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$	A ranking of items such that item π_i has a rank of i 5
$\boldsymbol{\pi^{-1}} = (\pi_1^{-1}, \ldots, \pi_n^{-1})$	Vector of ranks such that item j has a rank of π_j^{-1} 5
$\boldsymbol{w} = (w_1, \ldots, w_N)$	Scores attributed to each item 2
$\hat{\boldsymbol{w}} = (\hat{w}_1, \ldots, \hat{w}_N)$	Estimator of the scores attributed to each item 29
Y_i	Random variable describing the merit given to item i by a judge 8
$\boldsymbol{\theta}$	Additional parameters of the model; in the case of models allowing ties, $\boldsymbol{\theta} = \theta$ is the parameter describing the ties probability 17
$\hat{\boldsymbol{\theta}}$	Estimator of the additional parameters of the model; in the case of models allowing ties, $\hat{\boldsymbol{\theta}} = \hat{\theta}$ is the estimator of the parameter describing the ties probability 34
$p_{i,j,k}$ *(resp.* $p_{i,j}$*)*	Probability of observing the ranking i, j, k (resp. i, j) among the all the possible rankings 7
$p_{i=j}$	Probability of observing a tie between items i and j among the pair i, j 17
$p_{i,\cdot,\cdot}, p_{\cdot,i,\cdot}, p_{\cdot,i,\cdot}$ *(resp.* $p_{\cdot,i}, p_{i,\cdot}$*)*	Probability that item i has a rank 1, 2 or 3 (resp. 1 or 2) in a triplet i, j, k (resp. a pair i, j) of items 14
D_{ijk}	$= w_i^2(w_j + w_k) + w_j^2(w_i + w_k) + w_k^2(w_i + w_j)$, normalisation factor of the Mallows-Bradley-Terry model for triple comparisons 9
W_{ijk}	$= w_i + w_j + w_k$ 8
$k_{\boldsymbol{i,j,k}}$ *(reps.* $k_{\boldsymbol{i,j}}$*)*	Number of times the triplet i, j, k (resp. pair i, j) has been compared together 23

$k_{i,j,k}$ *(resp.* $k_{i,j}$*)*	Number of times the ranking i, j, k (resp. i, j) has been observed 23
$k_{i=j}$	Number of times a tie between items i and j has been observed among the pair i, j 24
K_i^l	Number of times item i takes position l in the ranking regarding all the comparisons done 23
$K^=$	Total number of ties among the observations, i.e. $K^= = \sum_{i=1}^{N} K_i^=$ 24
$h_{i,j,k}$ *(resp.* $h_{i,j}$*)*	Fraction of times the ranking i, j, k (resp. i, j) has been observed ($= k_{i,j,k}/k_{\boldsymbol{ijk}}$ (reps. $= k_{i,j}/k_{\boldsymbol{ij}}$)) 29
$h_{i=j}$	Fraction of times the tie between items i and j among the pair i, j has been observed ($= k_{i=j}/k_{\boldsymbol{ij}}$) 36
$\boldsymbol{R}$	Vector of ratio of h_*, the fraction of time a certain event occurs (for the BTL model, $R_{i,j} = h_{i,j}/h_{j,i}$) 29
$\boldsymbol{\rho}$	Vector of ratio of p_*, the probability that a certain event occurs (for the BTL model, $\rho_{i,j} = p_{i,j}/p_{j,i}$) 70
$\boldsymbol{Y}$	$= \log \boldsymbol{R}$ (element-wise logarithm of $\boldsymbol{R}$) 34
$\boldsymbol{\epsilon}$	Random variable representing the noise characterizing the observations $\boldsymbol{Y}$ 30
$\boldsymbol{z}$	Element-wise logarithm of the additional parameters and item scores, i.e. $\boldsymbol{z} = \left[(\log \boldsymbol{\theta})' \quad (\log \boldsymbol{w})'\right]'$ 30
$\hat{\boldsymbol{z}}$	Element-wise logarithm of the estimator of the additional parameters and item scores, i.e. $\hat{\boldsymbol{z}} = \left[\log(\hat{\boldsymbol{\theta}})' \quad (\log \hat{\boldsymbol{w}})'\right]'$ . . 30
V	Approximation of the covariance matrix of $\log \boldsymbol{R}$ 30
$\hat{V}$	Empirical approximation of the covariance matrix of $\log \boldsymbol{R}$ 31
D_V	Diagonal matrix with the approximated variance vector of $\log \boldsymbol{R}$ on the diagonal 63
$\hat{D}_V$	Diagonal matrix with the empirical approximated variance vector of $\log \boldsymbol{R}$ on the diagonal 63
n_{eq}	Number of equations composing the linear system in $\boldsymbol{z}$ . . 64
M'	Coefficients matrix corresponding to the equations of the studied model (of dimension $n_{eq} \times \tilde{N}$) 34
B'	Incidence matrix of the simple graph having the items as nodes and linking two items being compared together . . 30

S', m	Matrix and vector defining the additional constraints the parameters must satisfy in the case of non-observable parameters; in other words they define a restricted space $C = \{x \vert S'x = m\}$ 44
Q'	Orthonormal basis of $Ker(S')$ 44
$\boldsymbol{z^p}$	Particular vector $\boldsymbol{z}$ found by restricting the set $\boldsymbol{Y} = M'\boldsymbol{z} + \epsilon$; $\boldsymbol{z^p} = \left[\log(\boldsymbol{\theta^p})' \quad (\log \boldsymbol{w^p})'\right]'$ 68
$\hat{\boldsymbol{z}}_{\boldsymbol{c}}^{\boldsymbol{p}}$	BLUE of $\boldsymbol{z^p}$ satisfying $S'\hat{\boldsymbol{z}}_{\boldsymbol{c}}^{\boldsymbol{p}} = m$ (such that $S'\boldsymbol{z^p} = m$ as well) 70
$D_{\boldsymbol{w}}$	$= \text{diag}\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$ 73

Introduction

Evaluating item scores is a common problem which allows us to predict the result of a comparison or to obtain a global ranking among items (rank aggregation). Several everyday life activities require indeed to give some items more importance regarding others. Consider for example previous purchases of individuals on an e-commerce or the previous movies they watched on a streaming platform, one could be interested in the prediction of items/movies this subpopulation will like. Another common situation happens during sport tournament. Take for example a horse race, with the previous performances of the horses, one could want to have a global ranking of the different horse racers in order to predict which one will win the race.

This problem of estimating item scores that represent their strength/quality is not straightforward: finding explicit estimation of the scores is rarely feasible and asking users to assign an absolute score to items can be challenging because user scores might be poorly calibrated. Take for example the rating of a movie which only allow us to score it between 1 and 5. Different individuals also attribute scores differently, which might falsify the result. A film critic is indeed less likely to assign a maximum score to a movie while an individual who liked it might be more generous about its scoring.

One can ask a population to provide a global ranking of the items to directly recover a general user ranking of the items based on those observations. Several types of methods already exist for this problem, known as the rank aggregation problem [Lin10, FMM22]. Unfortunately, it can be complex for an individual to rank a large set of items, such as 300 movies for example, and global rankings are rarely available in practice, we will never observe a ranking of all horse racers.

Another way of recovering scores or a global ranking of the items, is to ask for (or to find) comparisons on a smaller set of items and then recovering scores or aggregated ranking. We will opt for this way since it is usually easier and more efficient to obtain regarding individual's global rankings. It would indeed be preferable to ask people (or to find) relative preferences in the form of item comparisons, instead of asking people to directly attribute a score to each proposed item or to rank all the items simultaneously. The comparisons observations can be performed by a judge, a person, a consumer, etc. In this report, we employed the term judge to denote the choice maker. Note that there is not always a judge, it is indeed possible to deduce item comparisons from common real-life observations. Choosing an item against others means indeed that someone prefers it compared to the rest, the result of a game can be a ranking or in the case of 2 players/teams, it can be understood as the winner is better ranked than the looser, watching a certain movie could reveal some information about its quality regarding other proposed movies.

Methods already exist for the recovering of a global ranking based on item compar-

isons, some of those methods allow to directly find a global ranking [SW17] and others allow us to find some item parameters that can be linked to their position in a global ranking. We consider a set of N items for which we aim to attribute a score $\boldsymbol{w}$ based on the observations. Several models linking item score parameters to the comparisons of those items exist. Using one model allows to obtain a particular intrinsic information about the merit of the items. The simplest comparison choice happens when two items are compared without the possibility of ties. The Bradley-Terry Luce model (BTL) [BT52a, Luc59] is famous and simple when dealing with those type of observations. It assumes that the probability that an item i is preferred against another j depends only on two positive intrinsic item parameters w_i and w_j and it is given by $w_i/(w_i + w_j)$.

The comparison outcome probability one gives can actually be a more complex function, which can depend on several parameters [Cat12]. First, it might depend on several *object-specific* parameters and not only one, the function might also be complex with unknown fixed parameters. A car comparison for example will probably depend on their design, their power, their safety, etc. Afterwards, the outcome might be *comparison-specific*. In a sport tournament, the home team is more likely to win against the opposite team as the players are familiar with the field. The outcome might also be influenced by the user with some *user-object-specific* parameters. In a university comparison, a student will indeed be more likely to choose a university where the courses are given in a language the student speaks, for example. Several applications and examples of those models can be found in [Cat12], such as a more detailed model overview.

All the previous parameters can be considered to predict the outcome of a future comparison or to find a global ranking, but the time might play a role as well. To attribute a merit to sport teams, one will indeed use more recent match results, as the results obtained 10 years ago are almost useless. The problem becomes indeed *dynamic* since those parameters might also vary with time and has been studied in recent years [BLSR20, KT21].

In the above we only considered pairwise comparisons where the outcome is not allowed to be a tie result, in other words, an item must be preferred against the other. In reality, those comparisons are in fact more difficult. We can indeed observe *ties* in the comparisons. In a sport game, two teams can obtain the same score at the end. It can also happen that an item is strongly or mild preferred to another item, which induces *preference degree* in the result of a comparison. Consider for example a sport team A winning $2-1$ against another team B and a team C who won $5-0$ against team B then preference degree is important.

Besides the pairwise comparisons, one can also consider comparisons of a *larger set of items*. Note that for a larger set of items, one can have access to observations of different form. If the set considered contains not as many items, it is still reasonable to ask for a ranking of all the items in the set, but even this information might not be available. It is indeed possible that the judges only provide the best item among the set of items, for example. In triple comparisons, we could also have observations of the worst item among the three, which would be equivalent to choose the two best items without knowing which one is ranked above the other one. The item scores can then be recovered again by using a model linking the observations to the scores as before.

The BTL model has already been widely studied, and various methods exist to recover the item weights based on comparison observations, namely the maximum likelihood, GSK (developed by Grizzle, Starmer and Koch), rank centrality and weighted least squares methods. Extensions of the simple pairwise comparison model are less present in the literature. Different models extending BTL model exist, and some methods were developed to recover the weights for some of them, among which the maximum likelihood method is very popular. The GSK method has also been extended for some pairwise comparison models with ties and triple comparison model. But the most recent method, namely the weighted least squares method, which gives nowadays the best results, has not already been extended for other comparison models.

Our aim is thus to extend the weighted least squares method developed for the BTL model for triple comparison models and pairwise comparison models with ties. It has been proven that this method has the best minimax optimal bounds (up to a scalar). We manage to prove that this method and the method developed for extensions of the BTL model is optimal as it belongs to a more general class of problems by assuming that the variance of the observations exists, namely find the Best Linear Unbiased Estimator of the parameters based on noisy observations from a linear system. For simple problems belonging to that class, theorems already exist to prove the optimality of the parameter estimator. However, for more complex problems, the Best Linear Unbiased Estimator of the parameters of the noisy linear system has not been found yet. Recovering item weights with a weighted least squares method is part of those complex problems, and we aim to extend existing theorems of BLUE to prove the optimality of the method.

This master thesis is organized as follows. Part I contains an introduction and discussion of the existing models extending the BTL model. This Part contains a summary of the literature, including the origins of the different models. In Part II, we present the existing methods for BTL and BTL extensions. The extension of the weighted least squares method we developed will be explained as well. The general problem will then be stated in Part III with optimality results and application to comparison models extending the BTL one. Part IV shows the empirical results obtained by the method we developed and compare the performances with the one of the existing methods. An application on attributing a score to Netflix movies and generating a global ranking is then presented.

Part I

Models

Based on ranking observations on a subset of n items from N items, we are interested in models defining probability functions. Take for example a set of $N = 20$ movies, the observations can be preference rankings on $n = 3$ movies, meaning that each judge will be asked to rank several triplets of movies. The aim is then to find a probabilistic model defining observation probabilities. More precisely, we define a ranking observation $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ as a permutation of the n items such that item π_i has a rank of i and $\boldsymbol{\pi^{-1}} = (\pi_1^{-1}, \ldots, \pi_n^{-1})$ as the vector of ranks such that item j has a rank of π_j^{-1}. Suppose that a judge response for the first three movies is $\boldsymbol{\pi} = (\text{movie } 2, \text{movie } 3, \text{movie } 1)$ then the rank of movie 3 is 2 ($\pi_2 = $ movie 3) and $\boldsymbol{\pi^{-1}} = (3, 1, 2)$.

We are interested in situations where people make discrete choice, which are realizations of random variable defined by continuous probabilistic models. In other words, the realization of the random variable can be a ranking of the proposed items $\boldsymbol{\pi}$, in which case we will study its probability $p(\boldsymbol{\pi})$ to occur. Realization of another variable is a relative rank of an object i regarding an object j, or even the specific rank of an object i, and we will be interested in the probabilities $p(\pi_i^{-1} < \pi_j^{-1})$ and $p(\pi_i^{-1})$ respectively. Several models have been developed in recent decades to define those probabilities, and we will focus on those considering that a quality/utility score w_i exists for each item i and defining probabilities based on their intrinsic vector $\boldsymbol{w}$. This means that probabilistic models, based on different score vectors $\boldsymbol{w}$, can provide similar observations coming from the corresponding probabilities.

The simplest ranking is for 2 items ($n = 2$) which can be seen as a pairwise comparison. The most basic model considers a probability $w_i/(w_i + w_j)$ of selecting the item i against j. It has been studied by Bradley and Terry [BT52a] and will be detailed in the next Chapter.

The literature contains different extensions for rankings of n items and their origins are summarized and explained in the next Sections: Plackett and Luce extension [Luc59, Pla75](Section 1.1) or Mallows, Bradley and Terry extension [Mal57] (Section 1.2). Those two models will be discussed in Section 1.3.Those models allow us to have probabilities for the observations based on triplets rankings [BT52b, PB60] for which 2 others models have been developed [Pen57] and briefly described in Section 1.4.

Some other types of extensions of the BTL model are presented in Chapter 2 to take ties [RK67, Dav70, BR73] and order effect [BG75] into account in the preferences. The ties extensions will be discussed in Section 2.1.

Chapter 1
BTL model and extensions to multi comparisons

The Bradley-Terry-Luce model [BT52a] is a statistic model that works only for pairwise comparisons. It supposes that for all pairs of objects that have been compared, the probability that one object i wins over the second j is given by:

$$p_{i,j} = \frac{w_i}{w_i + w_j} \tag{1.1}$$

where w_i is the real positive score attributed to the item i.

This model is very simple, as an object with a high weight will also have a high probability to win against another object. Following this model, if we ask an individual to choose between objects i and j, his answer will follow a Bernoulli probability law with a probability equal to $p_{i,j}$. In this model, it is mandatory that individuals choose between the two objects, they cannot choose both.

The next Sections will present the different models that extend this idea for rankings of three items. The BTL model has been extended for a ranking of any size n in two different ways: Plackett-Luce model and Mallow-Bradley-Terry model. Two other models extend the BTL model for rankings of three items : Freund model and first-second preferences triple model (those models could be extended for more than three, but this has not been done).

1.1 Plackett-Luce model (PL)

The model for triple comparison was obtained by Bradley & Terry [BT52b] and is described by

$$p_{i,j,k} = \frac{w_i w_j}{(w_i + w_j + w_k)(w_j + w_k)} \tag{1.2}$$

where the notation $p_{i,j,k} = p(\boldsymbol{\pi} = (\text{item } i, \text{item } j, \text{item } k))$ means the probability that object i wins over the two other objects and that object j wins over object k. As it was the case for the BTL model, in the PL model, the individuals are forced to choose a ranking between the proposed objects without any equality between items. An example using this model is given in Table 1.1. The Table entries correspond to the probabilities that a judge have to rank three movies in a certain order and results from movie's scores $\boldsymbol{w} = [3, 2, 1]$.

$p_{i,j,k}$	$p_{i,k,j}$	$p_{j,i,k}$	$p_{j,k,i}$	$p_{k,i,j}$	$p_{k,j,i}$
0.33	0.17	0.25	0.08	0.1	0.07

Table 1.1: Probabilities of choosing a specific ranking for the Plackett-Luce model for three items having scores equal to $\boldsymbol{w} = [3, 2, 1]$.

This triple comparison model (1.2) is a specific case of the Plackett-Luce model given by

$$p\left(\boldsymbol{\pi} = (\text{item } 1, \cdots, \text{item } n)\right) = \prod_{i=1}^{n-1} \frac{w_i}{\sum_{k=i}^{n} w_k}$$

for which, in the case of pairwise comparisons ($n = 2$), we get $p_{i,j} = \frac{w_i}{w_i + w_j}$ which is exactly the BTL model given by (1.1). For any ranking $\boldsymbol{\pi}$ this expression becomes

$$p\left(\boldsymbol{\pi}\right) = \prod_{i=1}^{n-1} \frac{w_{\pi_i}}{\sum_{k=i}^{n} w_{\pi_k}}$$

The Plackett-Luce model is based on the Luce's axiom that defines a probability of choosing item i among n items. In the case of two items, it is equal to $w_i/(w_i + w_j)$. For three items the probability becomes then $w_i/(w_i + w_j + w_k)$ and, more generally, $w_i/\sum_{j=1}^{n} w_j$ for n items.

The probability that the ranking i, j, k occurs is thus given by $p(\pi_1 = i)$, the probability of choosing the item i among all items and $p(\pi_2 = j|\pi_1 = i)$, the probability of choosing the item j among all items except item i (by denoting $W_{ijk} = w_i + w_j + w_k$):

$$\begin{aligned} p_{i,j,k} &= p(\pi_1 = i)p(\pi_2 = j|\pi_1 = i)p(\pi_3 = k|\pi_1 = i, \pi_2 = j) \\ &= \frac{w_i}{w_i + w_j + w_k} \frac{w_j}{w_j + w_k} \frac{w_k}{w_k} \\ &= \frac{w_i w_j}{W_{ijk}(w_j + w_k)} \end{aligned}$$

The Plackett-Luce model belongs in fact to a more general class of models, namely **Thurstone order statistics models**. A Thurstone order statistic model [Thu27] is a model such that an object i will have a merit μ_i. Each judge of the studied population gives a merit estimation Y_i with a probability function described by D_i and its cumulative distribution function F_i. What we are interested in is the probability that object i is preferred to object j by the individuals, which is denoted by $p_{i,j}$. As described in the equations below, this probability can be found by using a cumulative distribution function (cdf) F of the centered random variable $X_{ij} = (Y_j - \mu_j) - (Y_i - \mu_i)$ representing the difference of the estimated merit between items j and j:

$$p_{i,j} = p(Y_i > Y_j) = p\left(X_{ij} < (\mu_i - \mu_j)\right) = F_{X_{ij}}(\mu_i - \mu_j)$$

The **Plackett-Luce model** [Luc59, Pla75] is a Thurstone order statistics model with a Gumbel distribution function with the scale parameter equal to 1:

$$Y_i \sim \text{Gumbel}(\mu_i, 1)$$

This distribution has a cumulative distribution function given by

$$F_i(y, \mu_i) = \int_{-\infty}^{y} e^{-\left((t-\mu_i)+e^{(t-\mu_i)}\right)} dt = e^{-e^{-(y-\mu_i)}}$$

Furthermore, it is known that $(Y_j - \mu_j) - (Y_i - \mu_i)$ follows a standard logistic distribution:

$$Y_i - \mu_i \sim \text{Gumbel}(0,1) \implies (Y_j - \mu_j) - (Y_i - \mu_i) \sim \text{Logisitic}(0,1)$$

with a cumulative distribution function $F_{X_{ij}}(x) = \int_{-\infty}^{x} \frac{e^{-t}}{(1+e^{-t})^2} dt = \frac{1}{1+e^{-x}}$.

Denoting the positive item scores w_i by the change of variable e^{μ_i}, with that distribution law, it can be proven that [Luc59]:

$$p(Y_i > Y_j \ \forall j \in \{1, \cdots, n\} \backslash \{i\}) = \frac{w_i}{\sum_{k=1}^{n} w_k} = p(\pi_1 = i) \tag{1.3}$$

which is called the Luce's axiom. Following the same reasoning, we get:

$$p(Y_{i'} > Y_j \ \forall j \in \{1, \cdots, n\} \backslash \{i, i'\}) = \frac{w_{i'}}{\sum_{k=1 \neq i}^{n} w_k} = p(\pi_2 = i' | \pi_1 = i)$$

The probability of observing a particular ranking is given by the product of all those events:

$$p(\boldsymbol{\pi}) = \prod_{i=1}^{n-1} p(\pi_i | \pi_1, \ldots, \pi_{i-1}) = \prod_{i=1}^{n-1} \frac{w_{\pi_i}}{\sum_{k=i}^{n} w_{\pi_k}}$$

We observe that this model, first extension of the BTL model, allow us to perform a number of comparisons greater than 2 as it was the case for BTL. In this model, the ranking $\boldsymbol{\pi}$ that maximizes $p(\boldsymbol{\pi})$ will be the one that ranks the objects in the order of their scores w_i.

1.2 Mallows-Bradley-Terry model (MBT)

The model for triple comparisons has already been studied by Pendergrass & Bradley [PB60] and is defined by

$$p_{i,j,k} = \frac{w_i^2 w_j}{w_i^2 w_j + w_i^2 w_k + w_j^2 w_i + w_j^2 w_k + w_k^2 w_i + w_k^2 w_j} \tag{1.4}$$

where the denominator $D_{ijk} = w_i^2 w_j + w_i^2 w_k + w_j^2 w_i + w_j^2 w_k + w_k^2 w_i + w_k^2 w_j$ is the normalization term. An example using this model is given in Table 1.2. The same example as before is used, namely the Table entries correspond to the probabilities that a judge have to rank three movies in a certain order and results from movie's scores $\boldsymbol{w} = [3, 2, 1]$.

$p_{i,j,k}$	$p_{i,k,j}$	$p_{j,i,k}$	$p_{j,k,i}$	$p_{k,i,j}$	$p_{k,j,i}$
0.38	0.19	0.25	0.08	0.06	0.04

Table 1.2: Probabilities of choosing a specific ranking for the MBT model for three items having scores equal to $\boldsymbol{w} = [3, 2, 1]$.

The Pendergrass-Bradley model is a specific case of the Mallows-Bradley-Terry model given by

$$p\left(\boldsymbol{\pi} = (\text{item } 1, \cdots, \text{item } n)\right) = c(\boldsymbol{w}) \prod_{i=1}^{n-1} w_i^{n-i}$$

For any ranking $\boldsymbol{\pi}$, this expression becomes then

$$p\left(\boldsymbol{\pi}\right) = c(\boldsymbol{w}) \prod_{i=1}^{n-1} w_{\pi_i}^{n-i} \quad \text{with } c(\boldsymbol{w}) = \sum_{\boldsymbol{\pi} \in P} \prod_{i=1}^{n-1} w_{\pi_i}^{n-i}$$

where $c(\boldsymbol{w})$ is a scaling factor/normalization term and P is the set of all possible rankings of n items. For $n = 3$, the term $c(\boldsymbol{w})$ is equal to the inverse of $D_{ijk} = w_i^2 w_j + w_i^2 w_k + w_j^2 w_i + w_j^2 w_k + w_k^2 w_i + w_k^2 w_j$.

Again, in the case of pairwise comparisons ($n = 2$), we get $p_{i,j} = \frac{w_i}{w_i + w_j}$ which is exactly the BTL model given by (1.1).

The Mallows-Bradley-Terry model is based on the fact that a ranking is proportional to the BTL probabilities of each pair of items in the ranking. In the case of pairwise comparison, there is only one pair and $p_{i,j}$ was proportional to $w_i/(w_i + w_j)$. For three items, the probability $p_{i,j,k}$ becomes proportional to $p_{i,j}p_{i,k}p_{j,k}$ since there are three pairs in a set of three items. In other words, the probability that the ranking i, j, k occurs is proportional to the probabilities that i is ranked before j, j before k and i before k:

$$p_{i,j,k} \propto p_{i,j}p_{j,k}p_{i,k} = \frac{w_i}{w_i + w_j} \frac{w_j}{w_j + w_k} \frac{w_i}{w_i + w_k}$$

More generally, we have:

$$p(\boldsymbol{\pi}) \propto \prod_{1 \leq k < l \leq n} \frac{w_{\pi_k}}{w_{\pi_k} + w_{\pi_l}}$$

Note that all the probabilities sum to 1 thus this term is divided by the sum of the value for each possible ranking:

$$\begin{aligned} p(\boldsymbol{\pi}) &= c(\boldsymbol{p}) \prod_{1 \leq k < l \leq n} \frac{w_{\pi_k}}{w_{\pi_k} + w_{\pi_l}} \\ &= \left(\frac{\sum_{\boldsymbol{\pi} \in P} \prod_{1 \leq k < l \leq n} w_{\pi_k}}{\Pi} \right)^{-1} \frac{\prod_{1 \leq k < l \leq n} w_{\pi_k}}{\Pi} \\ &= c(\boldsymbol{w}) \prod_{1 \leq k < l \leq n} w_{\pi_k} \end{aligned}$$

It indeed holds that for any ranking $\boldsymbol{\pi}$ the expression $\Pi = \prod_{1\leq k<l\leq n}(w_{\pi_k} + w_{\pi_l})$ represents the product of the sums of all pairs one can recover from the n items and is thus always equal meaning that the normalization factor $c(\boldsymbol{p})$ can be written as

$$\sum_{\boldsymbol{\pi}\in P}\prod_{1\leq k<l\leq n}\frac{w_{\pi_k}}{w_{\pi_k}+w_{\pi_l}} = \frac{\sum_{\boldsymbol{\pi}\in P}\prod_{1\leq k<l\leq n} w_{\pi_k}}{\Pi}$$

The Mallow-Bradley-Terry belongs in fact to a more general class of models, namely the **Babington Smith models**. Those models are based on pairwise comparisons. In this model, the probability of observing a ranking $\boldsymbol{\pi}$ is given by:

$$p(\boldsymbol{\pi}) = c(\boldsymbol{p})\prod_{1\leq k<l\leq n} p_{\pi_k,\pi_l} \tag{1.5}$$

where the coefficient $c(\boldsymbol{p})$ represents the normalization as the sum of $p(\boldsymbol{\pi})$ over all possible rankings must be 1. If we let P be the set of all possible rankings, namely the set of all permutations of the n items, we get

$$c(\boldsymbol{p}) = \left(\sum_{\boldsymbol{\pi}\in P}\prod_{1\leq k<l\leq n} p_{\pi_k,\pi_l}\right)^{-1}$$

The **MBT model** [Mal57] is the second extension of the BTL model and is obtained by taking the BTL probabilities $p_{i,j} = \frac{w_i}{w_i+w_j}$ and replacing those values in (1.5) gives

$$p(\boldsymbol{\pi}) = c(\boldsymbol{w})\prod_{i=1}^{n-1} w_i^{n-\pi_i^{-1}}$$

1.3 Comparison of MBT and PL models

1.3.1 Similarities of the models

Note here that the PL model and the MBT model for $n = 2$ lead to the same model (BTL) but for $n \geq 3$, the two models are generally different as we can observe for $n = 3$:

$$p^{\mathrm{PL}}_{i,j,k} = \frac{w_i w_j}{(w_i+w_j+w_k)(w_j+w_k)}$$
$$p^{\mathrm{MBT}}_{i,j,k} = \frac{w_i^2 w_j}{w_i^2w_j + w_i^2w_k + w_j^2w_i + w_j^2w_k + w_k^2w_i + w_k^2w_j}$$

Case of equivalence of MBT and PL models One could wonder if for some values of $\boldsymbol{w}$, potentially different in the two models, the MBT model and the PL model are equivalent i.e. $p^{\mathrm{MBT}}_{i,j,k} = p^{\mathrm{PL}}_{i,j,k}$. This happens only when

$$\frac{\tilde{w}_i^2\tilde{w}_j}{\tilde{w}_i^2(\tilde{w}_j+\tilde{w}_k) + \tilde{w}_j^2(\tilde{w}_i+\tilde{w}_k) + \tilde{w}_k^2(\tilde{w}_i+\tilde{w}_j)} = \frac{w_i w_j}{(w_i+w_j+w_k)(w_j+w_k)}$$

for all the possible rankings (i, j, k). It results, after calculations, that the Mallows-Bradley-Terry model and the Plackett-Luce model reduce to the same model only when all the item scores $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}$ are identical (up to a multiplicative factor) and all equal: $\tilde{\boldsymbol{w}} = \alpha\boldsymbol{w}$ and $w_1 = w_2 = \cdots = w_N$. In this case, the probability of observing any ranking is 1/6.

1.3.2 Pairwise comparisons probabilities

In the context of the triple comparison models, one can also look in more details at the pairwise comparisons probabilities associated with a triplet ranking. Some of them were already developed in [BT52b] and [Pen57].

Conditional probabilities $p(\pi_i < \pi_j | \pi_k = \alpha)$ The probability that item i is ranked before item j in the ranking of the items i, j, k knowing that item k takes the position α is given by the Bayes theorem:

$$p(\pi_i < \pi_j | \pi_k = \alpha) = \frac{p(\pi_i < \pi_j \text{ and } \pi_k = \alpha)}{p(\pi_k = \alpha)} \tag{1.6}$$

From (1.6), $\boldsymbol{p(\pi_i < \pi_j | \pi_k = 1)}$ is equal to $p_{k,i,j}/(p_{k,i,j} + p_{k,j,i})$

Plackett-Luce

$$\begin{aligned} &= \frac{\frac{w_k}{W_{ijk}} \frac{w_i}{w_i+w_j}}{\frac{w_k}{W_{ijk}} \frac{w_i}{w_i+w_j} + \frac{w_k}{W_{ijk}} \frac{w_j}{w_i+w_j}} \\ &= \frac{w_i}{w_i + w_j} \end{aligned}$$

Mallows-Bradley-Terry

$$\begin{aligned} &= \frac{\frac{w_i w_k^2}{D_{ijk}}}{\frac{w_i w_k^2}{D_{ijk}} + \frac{w_j w_k^2}{D_{ijk}}} \\ &= \frac{w_i}{w_i + w_j} \end{aligned}$$

From (1.6), $\boldsymbol{p(\pi_i < \pi_j | \pi_k = 2)}$ is equal to $p_{i,k,j}/(p_{i,k,j} + p_{j,k,i})$

Plackett-Luce

$$\begin{aligned} &= \frac{\frac{w_i}{W_{ijk}} \frac{w_k}{w_j+w_k}}{\frac{w_i}{W_{ijk}} \frac{w_k}{w_j+w_k} + \frac{w_j}{W_{ijk}} \frac{w_k}{w_i+w_k}} \\ &= \frac{w_i(w_i + w_k)}{w_i(w_i + w_k) + w_j(w_j + w_k)} \end{aligned}$$

Mallows-Bradley-Terry

$$\begin{aligned} &= \frac{\frac{w_i^2 w_k}{D_{ijk}}}{\frac{w_i^2 w_k}{D_{ijk}} + \frac{w_j^2 w_k}{D_{ijk}}} \\ &= \frac{w_i^2}{w_i^2 + w_j^2} \end{aligned}$$

From (1.6), $\boldsymbol{p(\pi_i < \pi_j | \pi_k = 3)}$ is equal to $p_{i,j,k}/(p_{i,j,k} + p_{j,i,k})$

Plackett-Luce

$$\begin{aligned} &= \frac{\frac{w_i}{W_{ijk}} \frac{w_j}{w_j+w_k}}{\frac{w_i}{W_{ijk}} \frac{w_j}{w_j+w_k} + \frac{w_j}{W_{ijk}} \frac{w_i}{w_i+w_k}} \\ &= \frac{w_i + w_k}{w_i + w_j + 2w_k} \end{aligned}$$

Mallows-Bradley-Terry

$$\begin{aligned} &= \frac{\frac{w_i^2 w_j}{D_{ijk}}}{\frac{w_i^2 w_j}{D_{ijk}} + \frac{w_i w_j^2}{D_{ijk}}} \\ &= \frac{w_i}{w_i + w_j} \end{aligned}$$

We observe thus eventually that we recover the BTL model in both PL and MBT models for $p(\pi_i < \pi_j | \pi_k = 1)$ and only for the MBT model for $p(\pi_i < \pi_j | \pi_k = 3)$.

Taking the same example as mentioned above, Table 1.3 gives the different conditional probabilities among the three movies having scores $\boldsymbol{w} = [3, 2, 1]$. This example highlights again that the parameters $\boldsymbol{w}$ are intrinsic to the model meaning that they lead to different values of probability for the two considered models.

	$p(\pi_i < \pi_j \mid \pi_k)$			$p(\pi_i < \pi_k \mid \pi_j)$			$p(\pi_j < \pi_k \mid \pi_i)$		
	$\pi_k = 1$	$\pi_k = 2$	$\pi_k = 3$	$\pi_j = 1$	$\pi_j = 2$	$\pi_j = 3$	$\pi_i = 1$	$\pi_i = 2$	$\pi_i = 3$
MBT	0.60	0.69	0.60	0.75	0.90	0.75	0.67	0.80	0.66
PL	0.60	0.66	0.57	0.75	0.83	0.63	0.66	0.71	0.55

Table 1.3: Probabilities of preferring one item against another knowing the position of the third item in the ranking for the two studied triple comparison models and a set of three items having scores $\boldsymbol{w} = [3, 2, 1]$.

Probabilities $p(\pi_i < \pi_j)$ We can also compute $p_{i,j}$, the probability that in the triplet i, j, k the item i is ranked before item j, i.e. $p_{i,j,k} + p_{i,k,j} + p_{k,i,j}$.

For the Plackett-Luce model, as it was proven in [BT52b], this probability is simply given by the BTL probability. We indeed have:

$$\begin{aligned} p_{i,j} &= \frac{w_i w_j}{W_{ijk}(w_j + w_k)} + \frac{w_i w_k}{W_{ijk}(w_j + w_k)} + \frac{w_i w_k}{W_{ijk}(w_i + w_j)} \\ &= \frac{w_i}{W_{ijk}} \left(1 + \frac{w_k}{w_i + w_j}\right) = \frac{w_i}{w_i + w_j} \end{aligned}$$

The same reasoning was done for the Mallows-Bradley-Terry model in [Pen57]. We get in this case an expression different from the BTL model for pairwise comparisons.

$$p_{i,j} = \frac{w_i^2 w_j + w_i^2 w_k + w_k^2 w_i}{D_{ijk}}$$

For this probability, we get the BTL model for the PL model but not for the MBT model.

Taking the same example as mentioned above, Table 1.4 gives the different probabilities of rating a movie above another among the three movies having scores $\boldsymbol{w} = [3, 2, 1]$. The same observation about the fact that the parameter $\boldsymbol{w}$ is intrinsic to the model is made here. Pairwise comparisons based on the two considered models are indeed all different regarding the example chosen.

	$p_{i,j}$	$p_{i,k}$	$p_{j,k}$
MBT	0.625	0.813	0.708
PL	0.600	0.750	0.666

Table 1.4: Probabilities of rating an object above another in the triplet of items having scores $\boldsymbol{w} = [3, 2, 1]$ for the two triple comparisons models considered.

1.3.3 Rank probabilities

One may also be interested in the probability of choosing the rank of one item among the three items, denoted $p_{i,\cdot,\cdot}, p_{\cdot,i,\cdot}, p_{\cdot,\cdot,i}$ in function of the rank chosen for the

item i. Those probabilities are equal to the sum of the probabilities of the 2 rankings where the item i is ranked as chosen, i.e., $p_{i,j,k} + p_{i,k,j}, p_{j,i,k} + p_{k,i,j}$ and $p_{j,k,i} + p_{k,j,i}$ respectively.

Plackett-Luce In this model, those probabilities are computed hereunder, and it is equivalent to the Luce's axiom given by (1.3) in the first case (rank 1 for the item i):

$$
\begin{aligned}
p_{i,\cdot,\cdot} &= \frac{w_i w_j}{(w_i + w_j + w_k)(w_j + w_k)} + \frac{w_i w_k}{(w_i + w_j + w_k)(w_j + w_k)} = \frac{w_i}{w_i + w_j + w_k} \\
p_{\cdot,i,\cdot} &= \frac{w_i}{w_i + w_j + w_k} \frac{w_j^2 + w_k^2 + w_i(w_j + w_k)}{(w_i + w_j)(w_i + w_k)} \\
p_{\cdot,\cdot,i} &= \frac{w_j w_k}{w_i + w_j + w_k} \frac{w_j + w_k + 2w_i}{(w_i + w_j)(w_i + w_k)}
\end{aligned}
$$

Mallow-Bradley-Terry By the same reasoning, we get:

$$
\begin{aligned}
p_{i,\cdot,\cdot} &= \frac{w_i^2 w_j}{D_{ijk}} + \frac{w_i^2 w_k}{D_{ijk}} = \frac{w_i^2(w_j + w_k)}{w_i^2(w_j + w_k) + w_j^2(w_i + w_k) + w_k^2(w_i + w_j)} \\
p_{\cdot,i,\cdot} &= \frac{w_i(w_j^2 + w_k^2)}{w_i^2(w_j + w_k) + w_j^2(w_i + w_k) + w_k^2(w_i + w_j)} \\
p_{\cdot,\cdot,i} &= \frac{w_j w_k(w_j + w_k)}{w_i^2(w_j + w_k) + w_j^2(w_i + w_k) + w_k^2(w_i + w_j)}
\end{aligned}
$$

Numerical example: Taking the same example as mentioned above, Table 1.5 gives the different probabilities of attributing the position of one movie in the ranking of the three movies having scores $\boldsymbol{w} = [3, 2, 1]$.

	$p_{i,\cdot,\cdot}$	$p_{j,\cdot,\cdot}$	$p_{k,\cdot,\cdot}$	$p_{\cdot,i,\cdot}$	$p_{\cdot,j,\cdot}$	$p_{\cdot,k,\cdot}$	$p_{\cdot,\cdot,i}$	$p_{\cdot,\cdot,j}$	$p_{\cdot,\cdot,k}$
MBT	0.56	0.31	0.10	0.31	0.42	0.27	0.13	0.25	0.63
PL	0.5	0.33	0.17	0.35	0.40	0.25	0.15	0.27	0.58

Table 1.5: Probabilities of attributing the position of one item in the ranking of the triplet of items having scores $\boldsymbol{w} = [3, 2, 1]$ for the two triple comparisons models studied.

1.4 Other triple models

1.4.1 Freund Model

This model was proposed by Professor John E. Freund in [Pen57]. To obtain a ranking of three items i, j, k, we suppose that the objects are presented by pairs and that in any case, $Y_i > Y_j > Y_k$ where Y_i represents the merit estimation of item i by a judge. The pairs might be presented in 6 different ways, described below. The last column represents the probability of observing the ranking i, j, k.

First pair	Second pair	Third pair	
ij	ik	jk	$\frac{w_i}{w_i+w_j}\frac{w_i}{w_i+w_k}\frac{w_j}{w_j+w_k}$
ij	jk	ik	$\frac{w_i}{w_i+w_j}\frac{w_j}{w_j+w_k}$
ik	ij	jk	$\frac{w_i}{w_i+w_k}\frac{w_i}{w_i+w_j}\frac{w_j}{w_j+w_k}$
ik	jk	ij	$\frac{w_i}{w_i+w_k}\frac{w_j}{w_j+w_k}\frac{w_i}{w_i+w_j}$
jk	ik	ij	$\frac{w_j}{w_j+w_k}\frac{w_i}{w_i+w_k}\frac{w_i}{w_i+w_j}$
jk	ij	ik	$\frac{w_j}{w_j+w_k}\frac{w_i}{w_i+w_j}$

If the order of presentation of the pairs is random, the probability of observing the ranking i, j, k is given by the sum of the above expressions divided by six. The final expression is obtained by using some simple algebra:

$$p_{i,j,k} = \frac{w_i^2 w_j + w_i w_j w_k/3}{(w_i+w_j)(w_i+w_k)(w_j+w_k)}$$

This model is in fact not relevant in the study of models extending the BTL one. Indeed, by knowing the rank of one of the three items, we do not recover a BTL model.

1.4.2 First-second preferences triple model

This model given in [Pen57] is given by

$$p_{i,j,k} = \frac{\delta_i}{\delta_i+\delta_j+\delta_k}\frac{\gamma_j}{\gamma_j+\gamma_k}$$

It is composed of the second-preference parameters γ_i and of first-preference parameters δ_i such that the probability that an item i is preferred against items j and k is given by $\frac{\delta_i}{\delta_i+\delta_j+\delta_k}$ such as in the Luce's axiom. Those additional parameters make calculations unnecessarily more complex, and this model will thus not be further investigated. Note that this model would be equivalent to the Plackett-Luce model if the parameters $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ are equal. In this case, they would thus reduce to the parameters $\boldsymbol{w}$ of the PL model.

Chapter 2
Other extensions of BTL model

Models derived from BTL taking some other parameters than the scores $\boldsymbol{w}$ into account may be considered. This Chapter aims to present famous extensions of the BTL model, allowing judges to provide rankings with ties or to model the fact that the order of presentation of the items has its importance.

2.1 Models with ties

All the previous models assume that individuals are forced to choose items without any ties between two or three items. This limitation has been explored by Rao & Kupper and Davidson with MBT model, and two models (each of them for 2 and 3 items) have been developed to take ties into account. Note that, in the same way as the MBT model was extended for $n > 3$, those models could also be developed for larger item sets.

2.1.1 Rao & Kupper

Pairwise model Rao & Kupper [RK67] have generalized the BTL model by introducing a non-negative threshold $\nu = \ln \theta^{\mathrm{RK}}$ ($\theta^{\mathrm{RK}} > 1$) to reflect a judge's inability to choose between items even if there is a difference of scores, resulting in the following probabilities

$$p_{i,j} = \frac{w_i}{w_i + \theta^{\mathrm{RK}} w_j} \qquad\qquad p_{i=j} = \frac{((\theta^{\mathrm{RK}})^2 - 1) w_i w_j}{(w_i + \theta^{\mathrm{RK}} w_j)(\theta^{\mathrm{RK}} w_i + w_j)}$$

Remember from the Chapter 1 that the BTL model is a particular case of Thurstone order statistics models, in other words, an item i will have a merit μ_i and each judge will give a merit estimation Y_i to the item i. The probability that item i is ranked above item j is then given by $p_{i,j} = p\left(Y_i > Y_j\right) = p\left(X_{ij} < (\mu_i - \mu_j)\right)$ for $X_{ij} = (Y_j - \mu_j) - (Y_i - \mu_i)$. For the pairwise comparison model, X_{ij} follows a Logistic distribution with a location parameter equal to 0 and a unit scale parameter. Furthermore, the item scores are such that $w_i = e^{\mu_i}$.

The Rao-Kupper model assumes that items i and j have the same merit when $Y_j - Y_i$ lies below the threshold:

$$p_{i,j} = p\left(Y_j - Y_i < -\nu\right) \qquad p_{i=j} = p\left(\left|Y_j - Y_i\right| < \nu\right) \qquad p_{j,i} = p\left(Y_j - Y_i > \nu\right)$$

From the standard Logistic distribution, it holds that $p(X < x) = (e^{-x} + 1)^{-1}$. The above expression becomes then

$$p_{i,j} = p\left(Y_j - Y_i < -\nu\right) = \frac{1}{e^{-(\mu_i - \mu_j - \nu)} + 1} = \frac{w_i}{w_i + \theta^{\mathrm{RK}} w_j}$$
$$p_{i=j} = p\left(\left|Y_j - Y_i\right| < \nu\right) = \frac{1}{e^{-(\mu_i - \mu_j + \nu)} + 1} - \frac{1}{e^{-(\mu_i - \mu_j - \nu)} + 1} = \frac{((\theta^{\mathrm{RK}})^2 - 1) w_i w_j}{(w_i + \theta^{\mathrm{RK}} w_j)(\theta^{\mathrm{RK}} w_i + w_j)}$$
$$p_{j,i} = p\left(Y_j - Y_i > \nu\right) = 1 - \frac{1}{e^{-(\mu_i - \mu_j + \nu)} + 1} = \frac{w_j}{\theta^{\mathrm{RK}} w_i + w_j}$$

An example using this model is given in Table 2.1. The Table entries corresponds to the probabilities that a judge ranks two movies in a certain order with possibility of ties. Those probabilities result from movie's scores $\boldsymbol{w} = [2, 3]$ and a threshold $\theta^{\mathrm{RK}} = 1.5$.

$p_{i,j}$	$p_{i=j}$	$p_{j,i}$
0.31	0.19	0.5

Table 2.1: Probabilities of choosing a specific ranking for the Rao-Kupper model for two items having score equal to $\boldsymbol{w} = [2, 3]$ and a threshold $\theta^{\mathrm{RK}} = 1.5$.

Triple model The previous model has been extended to three items in the same way as the MBT models by Beaver and Rao [BR73] to allow ties between two or three items in triple comparisons. The new probabilities are described by

$$p_{i,j,k} = \frac{w_i^2 w_j D_{ji} D_{ki} D_{kj}}{A_{ijk}} \qquad p_{i,j=k} = \frac{w_i^2 w_j w_k D_{ji} D_{ki} ((\theta^{\mathrm{RK}})^2 - 1)}{A_{ijk}}$$
$$p_{i=j,k} = \frac{w_i^2 w_j^2 D_{ki} D_{kj} ((\theta^{\mathrm{RK}})^2 - 1)}{A_{ijk}} \qquad p_{i=j=k} = \frac{w_i^2 w_j^2 w_k^2 ((\theta^{\mathrm{RK}})^2 - 1)^3}{A_{ijk}}$$

with $D_{ij} = w_i + \theta^{\mathrm{RK}} w_j$ and A_{ijk}, the normalization term:

$$A_{ijk} = \sum_P w_r^2 w_s D_{sr} D_{tr} D_{ts} + ((\theta^{\mathrm{RK}})^2 - 1) \sum_P w_r^2 w_s w_t D_{sr} D_{tr} + ((\theta^{\mathrm{RK}})^2 - 1)^3 w_i^2 w_j^2 w_k^2$$

where P is the set of the six permutations of the triplet i, j, k.

In those models, when $\theta^{\mathrm{RK}} = 1$ ($\nu = 0$) the ties probabilities become zero and the probabilities are the same as in the original BTL (1.1) and MBT model (1.4).

2.1.2 Davidson

Pairwise model Davidson [Dav70] introduced a model extending the BTL model where the probability of non-preference is proportional to the geometric mean of the probabilities of preferences, i.e. $\sqrt{p_{i,j} p_{j,i}}$:

$$p_{i,j} = \frac{w_i}{w_i + w_j + \theta^{\mathrm{D}} \sqrt{w_i w_j}} \qquad p_{i=j} = \frac{\theta^{\mathrm{D}} \sqrt{w_i w_j}}{w_i + w_j + \theta^{\mathrm{D}} \sqrt{w_i w_j}}$$

This model is based on the Luce axiom, which states that the probability of choosing an item i among n items is $w_i / \sum_{j=1}^{n} w_j$. A direct consequence of that axiom is that

$p_{i,j}/p_{j,i} = w_i/w_j$. The BTL model is recovered by knowing that $p_{i,j} + p_{j,i} = 1$. The Davidson model aims to keep the above property while allowing ties in the comparisons of items. To do so, the probability of observing a tie between two items i and j is said to be proportional to the geometric mean of the probabilities of preference for the pair of items considered:

$$p_{i=j} \propto \sqrt{p_{i,j}p_{j,i}}$$

As it must hold that $p_{i,j} + p_{j,i} + p_{i=j} = 1$ and using the consequence of the Luce axiom, we obtain

$$p_{i,j} = \left(\frac{p_{i,j}}{p_{i,j}} + \frac{p_{j,i}}{p_{i,j}} + \frac{p_{i=j}}{p_{i,j}}\right)^{-1} = \left(1 + \frac{w_j}{w_i} + \frac{\theta^{\mathrm{D}}\sqrt{w_i w_j}}{w_i}\right)^{-1} = \frac{w_i}{w_i + w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}}$$

$$p_{i=j} = \left(\frac{p_{i,j}}{p_{i=j}} + \frac{p_{j,i}}{p_{i=j}} + \frac{p_{i=j}}{p_{i=j}}\right)^{-1} = \left(\frac{w_i}{\theta^{\mathrm{D}}\sqrt{w_i w_j}} + \frac{w_j}{\theta^{\mathrm{D}}\sqrt{w_i w_j}} + 1\right)^{-1} = \frac{\theta^{\mathrm{D}}\sqrt{w_i w_j}}{w_i + w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}}$$

$$p_{j,i} = \left(\frac{p_{i,j}}{p_{j,i}} + \frac{p_{j,i}}{p_{j,i}} + \frac{p_{i=j}}{p_{j,i}}\right)^{-1} = \left(\frac{w_i}{w_j} + 1 + \frac{\theta^{\mathrm{D}}\sqrt{w_i w_j}}{w_j}\right)^{-1} = \frac{w_j}{w_i + w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}}$$

An example using this model is given in Table 2.2. The same example as before is used, namely the Table entries corresponds to the probabilities that a judge ranks two movies in a certain order with possibility of ties. Those probabilities result from movie scores $\boldsymbol{w} = [2, 3]$ and a threshold parameter equal to $\theta^{\mathrm{D}} = 0.5$.

$p_{i,j}$	$p_{i=j}$	$p_{j,i}$
0.32	0.20	0.48

Table 2.2: Probabilities of choosing a specific ranking for the Davidson model for two items having scores equal to $\boldsymbol{w} = [2, 3]$ and a threshold $\theta^{\mathrm{D}} = 0.5$.

Triple model Again, the Davidson model has been extended by Beaver and Rao [BR73] for triple comparisons like the MBT model based on the pairwise comparison model. The probabilities are given by

$$p_{i,j,k} = \frac{w_i^2 w_j}{G_{ijk}} \qquad p_{i,j=k} = \frac{\theta^{\mathrm{D}} w_i^2 \sqrt{w_j w_k}}{G_{ijk}}$$

$$p_{i=j,k} = \frac{\theta^{\mathrm{D}} w_i w_j \sqrt{w_i w_j}}{G_{ijk}} \qquad p_{i=j=k} = \frac{(\theta^{\mathrm{D}})^3 w_i w_j w_k}{G_{ijk}}$$

with $G_{ijk} = \sum_P w_s^2 w_t + \frac{\theta^{\mathrm{D}}}{2}\sum_P \sqrt{w_s^3 w_t^3} + \sum_P \sqrt{w_r^4 w_s w_t} + (\theta^{\mathrm{D}})^3 w_i w_j w_k$ the normalization term.

In those models, when $\theta^{\mathrm{D}} = 0$ the ties probabilities become zero and the probabilities are the same as in the original BTL (1.1) and MBT model (1.4).

2.1.3 Case of equivalence of Rao & Kupper and Davidson models

As it was mentioned in the previous Section, the two considered models are equivalent to a BTL model when the additional parameter θ is equal to 1 and 0 for the

Rao-Kupper and the Davidson model respectively (in other words, when the parameter is chosen such that the probability to observe ties is null).

One could also wonder when the two models are equivalent in a more general point of view. This happens when the different probabilities are equal:

$$p_{i,j}^{\mathrm{RK}} = p_{i,j}^{\mathrm{D}} \ \forall i,j \in \{1,\cdots,N\} \times \{1,\cdots,N\}$$

This holds only when scores parameters $\boldsymbol{w}$ are the same in both models, $\boldsymbol{w}^{\mathrm{RK}} = \boldsymbol{w}^{\mathrm{D}}$, ties parameter satisfy $\theta^{\mathrm{RK}} = \theta^{\mathrm{D}} + 1$ and all the scores are identical $w_i = w_j \ \forall i,j$. In this case,

$$p_{i,j} = \frac{1}{1+\theta^{\mathrm{RK}}} = \frac{1}{2+\theta^{\mathrm{D}}} \qquad\qquad p_{i=j} = \frac{\theta^{\mathrm{RK}}-1}{1+\theta^{\mathrm{RK}}} = \frac{\theta^{\mathrm{D}}}{2+\theta^{\mathrm{D}}}$$

2.1.4 Non-preference probabilities

It is interesting in the cases of comparison models with ties to compute the non-preference probability. Assume that we are interested in the probability that an item i is not strictly ranked above item j, this probability is equal to $p_{i=j} + p_{j,i} = 1 - p_{i,j}$ and it will be developed below for the two models studied.

Rao-Kupper model $\quad 1 - p_{i,j} = 1 - \dfrac{w_i}{w_i + \theta^{\mathrm{RK}} w_j} = \dfrac{\theta^{\mathrm{RK}} w_j}{w_i + \theta^{\mathrm{RK}} w_j}$

Davidson model $\quad 1 - p_{i,j} = 1 - \dfrac{w_i}{w_i + w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}} = \dfrac{w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}}{w_i + w_j + \theta^{\mathrm{D}}\sqrt{w_i w_j}}$

Taking the same example as mentioned above, Table 2.3 gives the different non-preference probabilities for the two movies having scores $\boldsymbol{w} = [2,3]$ and a threshold parameter equal to $\theta^{\mathrm{RK}} = 1.5$ and $\theta^{\mathrm{D}} = 0.5$.

	$1 - p_{i,j}$	$1 - p_{j,i}$
Rao-Kupper	0.69	0.5
Davidson	0.68	0.42

Table 2.3: Non-preference probabilities of the items for the two pairwise comparison models with ties studied for two items having score equal to $\boldsymbol{w} = [2,3]$ and a threshold $\theta^{\mathrm{RK}} = 1.5$ and $\theta^{\mathrm{D}} = 0.5$.

2.2 Models with order affect

Beaver and Gokhale [BG75] have proposed in 1975 a generalization of the BTL model to account the order of presentation of the two items. They introduce an additive term θ_{ij} and the parameters w_i and w_j of the BTL model are replaced by $w_i + \theta_{ij}$ and $w_j - \theta_{ij}$ in the probabilities

$$p_{i,j} = \frac{w_i + \theta_{ij}}{w_i + w_j} \qquad p_{j,i} = \frac{w_j - \theta_{ij}}{w_i + w_j} \qquad |\theta_{ij}| \leq \min(w_i, w_j)$$

Part II

Methods

Individual preferences can be expressed based on the scores $\boldsymbol{w}$ thanks to the models introduced in the previous Part. Methods have been developed to recover the parameters w_i of the model considered, based on observations of the preferences and the hypothesis that preferences follow a certain model. Those methods will be explained in Chapter 3.

The oldest introduced method to recover the scores is the one presented by the authors of the several models and is based on maximum log-likelihood of the probability of observing those preferences. Section 3.1 details how it has been applied to the BTL model [BT52a, Hun04], to the two BTL models with ties (Rao-Kupper [RK67] and Davidson [Dav70]) and to the two main models for triplets (PL [Hun04] and MBT [PB60]).

The second method [Bea77], presented in Section 3.2, similar to the one we developed, follows an weighted least squares approach and has been used for the BTL, Rao-Kupper, Davidson and MBT models in 1977.

More recently, two other approaches have been presented to recover the weights of the BTL model: the rank centrality method [NOS12] employing a random walk (Section 3.3) and a new weighted least squares approach [HOS20]. This second approach is detailed in Chapter 3.4 and includes different methods.

This last approach showed the best performance results for the BTL model, and we extended it for the other models presented and discussed in Part I hoping to reach similar performances. Chapter 4 presents thus the extension of those methods we developed for the two BTL models with ties and the two main triplet comparison models.

Chapter 3
Existing methods

This Chapter aims to describe the methods currently used to recover item weights based on judges' comparisons. Among them, we find the maximum log-likelihood, the GSK, the rank centrality and the recent weighted least squares methods explained in Sections 3.1, 3.2, 3.3 and 3.4 respectively.

3.1 Maximum likelihood estimation methods

This method aims to find the vector of scores $\boldsymbol{w}$ for which the probability of observing the global ranking associated to that vector is the highest. To that purpose, we need to solve $\partial l(\boldsymbol{w})/\partial w_l = 0$ where $l(\boldsymbol{w})$ is the logarithm of the probability of observing the ranking associated to $\boldsymbol{w}$. As we will see, there is a model dependent function f for which solving the equation is equivalent to find $\boldsymbol{w}$ satisfying $\boldsymbol{w} = f(\boldsymbol{w})$. The latter expression can be solved using a fixed point iteration method, which consists of updating the vector of weights the following way: $\boldsymbol{w}^{t+1} = f(\boldsymbol{w}^t)$ starting from a prior vector $\boldsymbol{w}^0$.

3.1.1 Bradley-Terry-Luce model

The maximum (log-) likelihood of $\boldsymbol{w}$ was derived by Bradley and Terry in [BT52a] and further developed in [Hun04]. If $k_{i,j}$ represents the number of times item i is chosen against j and using the BTL model we get the log-likelihood function

$$l(\boldsymbol{w}) = \ln\left(\prod_{i=1}^{N}\prod_{j=1}^{N} p_{i,j}^{k_{i,j}}\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}\Big(k_{i,j}\ln w_i - k_{i,j}\ln(w_i + w_j)\Big)$$

We are looking for the values of $\boldsymbol{w}$ that maximize $l(\boldsymbol{w})$. Defining $k_{\boldsymbol{ij}}$ as the total number of comparisons performed between items i and j ($k_{\boldsymbol{ij}} = k_{i,j} + k_{j,i}$) we get:

$$\frac{\partial l(\boldsymbol{w})}{\partial w_l} = \sum_{\substack{j=1\\ j\neq l}}^{N}\frac{k_{l,j}}{w_l} - \frac{k_{\boldsymbol{jl}}}{w_l + w_j} = 0$$

We can thus find the values of w_i for the different items using an iterative process:

$$w_i^{t+1} = K_i^1\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\frac{k_{\boldsymbol{ij}}}{w_i^t + w_j^t}\right)^{-1}$$

where $K_i^1 = \sum_{j=1}^{N} k_{i,j}$ denotes the number of times item i wins against all other items.

3.1.2 Rao & Kupper model

The log-likelihood function for this model is given by

$$
\begin{aligned}
l(\theta, \boldsymbol{w}) &= \ln \left(\prod_{(i,j)\in P} p_{i,j}^{k_{i,j}} p_{j,i}^{k_{j,i}} p_{i=j}^{k_{i=j}} \right) \\
&= \sum_{(i,j)\in P} \Big(k_{i,j} \left(\ln w_i - \ln(w_i + \theta w_j) \right) + k_{j,i} \left(\ln w_j - \ln(\theta w_i + w_j) \right) \\
&\qquad + k_{i=j} \left(\ln(\theta^2 - 1) + \ln w_i + \ln w_j - \ln(w_i + \theta w_j) - \ln(\theta w_i + w_j) \right) \Big)
\end{aligned}
$$

where P denotes the set of pairs of items being compared to one another and $k_{i=j}$ is the number of ties observations of between the items i and j. We are again looking for the item weights $\boldsymbol{w}$ maximizing $l(\theta, \boldsymbol{w})$:

$$
\frac{\partial l(\theta, \boldsymbol{w})}{\partial w_l} = \sum_{\substack{j=1 \\ j\neq l}}^{N} \frac{k_{l,j} + k_{l=j}}{w_l} - \frac{k_{l,j} + k_{l=j}}{w_l + \theta w_j} - \theta \frac{k_{j,l} + k_{l=j}}{\theta w_l + w_j} = 0
$$

Note in the above expression that to recover the item weights, one need to estimate the additional parameter θ as well. In a similar way as above, we obtain:

$$
\frac{\partial l(\theta, \boldsymbol{w})}{\partial \theta} = \sum_{(i,j)\in P} -\frac{w_j(k_{i,j} + k_{i=j})}{w_i + \theta w_j} - \frac{w_i(k_{j,i} + k_{i=j})}{\theta w_i + w_j} + \theta \frac{2k_{i=j}}{\theta^2 - 1} = 0
$$

We can again find the values of w_i for the different items using an iterative process:

$$
\begin{aligned}
w_i^{t+1} &= (K_i^1 + K_i^=) \left(\sum_{\substack{j=1 \\ j\neq i}}^{N} \frac{k_{i,j} + k_{i=j}}{w_i^t + \theta^t w_j^t} + \theta^t \frac{k_{j,i} + k_{i=j}}{\theta^t w_i^t + w_j^t} \right)^{-1} \\
\theta^{t+1} &= 2K^= \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{w_i^t (k_{j,i} + k_{i=j})}{\theta^t w_i^t + w_j^t} \right)^{-1} + \frac{1}{\theta^t}
\end{aligned}
$$

where K_i^1 and $K_i^= = \sum_{j=1}^{N} k_{i=j}$ denote the number of times item i wins against all other items and the number of times item i is tied with all other items respectively and $K^= = \sum_{i=1}^{N} \sum_{j=i+1}^{N} k_{i=j}$ is the total number of ties among the observations.

3.1.3 Davidson model

For the Davidson model, the log-likelihood function is given by

$$
\begin{aligned}
l(\theta, \boldsymbol{w}) &= \ln \left(\prod_{(i,j)\in P} p_{i,j}^{k_{i,j}} p_{j,i}^{k_{j,i}} p_{i=j}^{k_{i=j}} \right) \\
&= \sum_{(i,j)\in P} \Big(k_{i,j} \left(\ln w_i - \ln(w_i + w_j + \theta\sqrt{w_i w_j}) \right) + k_{j,i} \left(\ln w_j - \ln(w_i + w_j + \theta\sqrt{w_i w_j}) \right) \\
&\qquad + k_{i=j} \left(\ln\theta + \frac{1}{2}\ln w_i + \frac{1}{2}\ln w_j - \ln(w_i + \theta w_j) - \ln(w_i + w_j + \theta\sqrt{w_i w_j}) \right) \Big)
\end{aligned}
$$

With the same reasoning as for the Rao & Kupper model, one obtains the following iterative method:

$$w_i^{t+1} = \left(K_i^1 + \frac{K_i^=}{2}\right)\left(\sum_{\substack{j=1\\ j\neq i}}^{N} \frac{k_{\boldsymbol{ij}}\left(1+\frac{\theta^t}{2}\sqrt{\frac{w_j^t}{w_i^t}}\right)}{w_i^t + w_j^t + \theta^t\sqrt{w_i^t w_j^t}}\right)^{-1}$$

$$\theta^{t+1} = K^= \left(\sum_{i=1}^{N}\sum_{j=1}^{N} \frac{1}{2}\frac{k_{\boldsymbol{ij}}\sqrt{w_i^t w_j^t}}{w_i^t + w_j^t + \theta^t\sqrt{w_i^t w_j^t}}\right)^{-1}$$

where $k_{\boldsymbol{ij}} = k_{i,j} + k_{j,i} + k_{i=j}$ in this case.

3.1.4 Plackett-Luce model

We can apply the same methodology for the Plackett-Luce model. The log-likelihood function is given below, where $k_{i,j,k}$ represents the number of times the ranking i, j, k is chosen among the six possible rankings.

$$\begin{aligned} l(\boldsymbol{w}) &= \ln\left(\prod_{i=1}^{N}\prod_{j=1}^{N}\prod_{k=1}^{N} p_{i,j,k}^{k_{i,j,k}}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\Big(k_{i,j,k}\ln w_i + k_{i,j,k}\ln w_j - k_{i,j,k}\ln(W_{ijk}) - k_{i,j,k}\ln(w_j+w_k)\Big) \end{aligned}$$

The maximum of this function can then be found by applying the following iterative method:

$$w_i^{t+1} = (K_i^1 + K_i^2)\left(\sum_{\substack{j=1\\ j\neq i}}^{N}\sum_{\substack{k=1\\ k\neq i}}^{N} \frac{k_{i,j,k}+k_{j,i,k}+k_{j,k,i}}{w_i^t + w_j^t + w_k^t} + \frac{k_{j,i,k}+k_{j,k,i}}{w_i^t + w_k^t}\right)^{-1}$$

where K_i^l denote the number of times the item i takes position l in the ranking regarding all the comparisons done. This result was obtained in [Hun04].

3.1.5 Mallows-Bradley-Terry model

The same computations can be applied for the Mallows-Bradley-Terry model for triple comparisons, as shown below:

$$\begin{aligned} l(\boldsymbol{w}) &= \ln\left(\prod_{i=1}^{N}\prod_{j=1}^{N}\prod_{k=1}^{N} p_{i,j,k}^{k_{i,j,k}}\right) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N}\Big(2k_{i,j,k}\ln w_i + k_{i,j,k}\ln w_j - k_{i,j,k}\ln(D_{ijk})\Big) \end{aligned}$$

The iterative process for recovering the item scores is then

$$w_i^{t+1} = \frac{(2K_i^1 + K_i^2)}{\sum\limits_{\substack{j=1\\ j\neq i}}^{N}\sum\limits_{\substack{k=1\\ k\neq i}}^{N} \frac{2w_i^t(w_j^t+w_k^t)+(w_j^t)^2+(w_k^t)^2}{(w_i^t)^2(w_j^t+w_k^t)+(w_j^t)^2(w_i^t+w_k^t)+(w_k^t)^2(w_i^t+w_j^t)}(k_{i,j,k}+k_{j,i,k}+k_{j,k,i})}$$

This result was obtained in [PB60].

3.2 GSK method

Grizzle, Starmer, and Koch [GSK69] have presented a method to estimate parameters of a linear system that define functions of categorical data probabilities. Let $\boldsymbol{p}$ be the vector of probabilities $\boldsymbol{p}' = \begin{pmatrix} \boldsymbol{p}'_1 & \boldsymbol{p}'_2 & \dots & \boldsymbol{p}'_{n_s} \end{pmatrix}$ with n_s the number of subsets of items, e.g., the number of pairs or triplets, and $\boldsymbol{p}'_i = \begin{pmatrix} p_{i1} & p_{i2} & \dots & p_{ir} \end{pmatrix}$, the vector of probabilities such that $\sum_{j=1}^{r} p_{ij} = 1$ with r the number of categories/outcomes, e.g., 2 (possible choices) for pairs, 3 for pairs with ties and 6 (possible rankings) for triplets without ties. If a, b is the i^{th} pair of item in a pairwise comparison model, it holds thus that $p_{i1} = p_{a,b}$ (resp. $p_{i2} = p_{b,a}$) is the probability that item a (resp. b) is preferred against item b (resp. a).

Let us introduce:

- $\boldsymbol{h}$, the sample estimate of the vector of probabilities $\boldsymbol{p}$
- $\hat{V}(\boldsymbol{h})$, the block diagonal matrix with the blocks $\hat{V}(\boldsymbol{h_i})$, the sample estimate of the covariance matrix of $\boldsymbol{h_i}$
- u functions of the vector $\boldsymbol{p}$, $f_m(\boldsymbol{p})$ and $[F(\boldsymbol{p})]' = \begin{bmatrix} f_1(\boldsymbol{p}) & f_2(\boldsymbol{p}) & \dots & f_u(\boldsymbol{p}) \end{bmatrix}$
- $H = \left[\dfrac{\partial f_m(\boldsymbol{p})}{\partial p_{ij}} \middle| \boldsymbol{p} = \boldsymbol{h} \right]$ of dimension $x \times rn_s$
- $S = H\hat{V}(\boldsymbol{h})H'$ of dimension $u \times u$

There are 2 special cases: when $F(\boldsymbol{p}) = A\boldsymbol{p}$ then $H = A$ or when $F(\boldsymbol{p}) = K \log(A\boldsymbol{p})$ then $H = KD_a^{-1}A$ where D_a is the diagonal matrix with the elements of $A\boldsymbol{p}$ on the diagonal.

The system to approximate will be $F(\boldsymbol{p}) = X\boldsymbol{y} \approx F(\boldsymbol{h})$ where X is a known matrix and $\boldsymbol{y}$ the unknown vector of parameters. The estimate of $\boldsymbol{y}$ by weighted least squares is $\hat{\boldsymbol{y}} = (X'S^{-1}X)^{-1}(X'S^{-1})F$.

Beaver [Bea77] applied this approach to estimate the weights w_i of the BTL model, pairwise comparison with ties models and triple comparison MBT model.

3.2.1 Bradley-Terry-Luce pairwise model

In the case of pairwise comparisons among N items, the number of categories, r, is 2 and $n_s = \frac{N(N-1)}{2}$, such that:

$$\boldsymbol{p}' = \begin{pmatrix} p_{1,2} & p_{2,1} & p_{1,3} & p_{3,1} & \dots & p_{N,N-1} \end{pmatrix} \qquad \text{where} \qquad p_{i,j} = \frac{w_i}{w_i + w_j}$$

Let $h_{i,j}$, the estimate of $p_{i,j}$, be equal to $\frac{k_{i,j}}{k_{ij}}$ with $k_{i,j}$ the number of preferences for item i against item j and k_{ij} be the number of comparisons involving items i and j. If

$k_{i,j} = 0$, replace $k_{i,j}$ by $\frac{1}{2k_{ij}}$ and $k_{j,i} = 1 - k_{i,j}$

Define $f_{ij}(\boldsymbol{p}) = \log\left(\frac{p_{i,j}}{p_{j,i}}\right)$, $f_{ij}(\boldsymbol{h}) = f_{ij}(\boldsymbol{p}|\boldsymbol{p} = \boldsymbol{h})$ such that:

$$F(\boldsymbol{p}) = K\log(\boldsymbol{p}) = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & -1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -1 \end{bmatrix} \log\left(\begin{bmatrix} p_{1,2} \\ p_{2,1} \\ p_{1,3} \\ p_{3,1} \\ \vdots \\ p_{N-1,N} \\ p_{N,N-1} \end{bmatrix}\right)$$

Let us introduce $y_i = \log w_i - \log w_N$, such that:

$$F(\boldsymbol{p}) = X\boldsymbol{y} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{N-1} \end{bmatrix}$$

Estimates of the w_i can be obtained from the $\hat{y}_i$ by $\hat{w_i} = \frac{e^{\hat{y_i}}}{\sum_{j=1}^{N} e^{\hat{y_j}}}$.

3.2.2 Pairwise models with ties

In the case of pairwise comparisons with ties among N items, the number of categories, r, is 3 and $n_s = \frac{N(N-1)}{2}$, such that:

$$\boldsymbol{p}' = \begin{pmatrix} \boldsymbol{p'_{1,2}} & \boldsymbol{p'_{1,3}} & \dots & \boldsymbol{p'_{N-1,N}} \end{pmatrix} \qquad \text{where} \qquad \boldsymbol{p'_{i,j}} = \begin{pmatrix} p_{i,j} & p_{j,i} & p_{i=j} \end{pmatrix}$$

Rao-Kupper model has the following probabilities:

$$p_{i,j} = \frac{w_i}{w_i + \theta w_j} \qquad\qquad p_{i=j} = \frac{(\theta^2 - 1)w_i w_j}{(w_i + \theta w_j)(\theta w_i + w_j)}$$

Let us define $\boldsymbol{f_{ij}}(\boldsymbol{p})$, A and K such that $F(\boldsymbol{p}) = K\log(A\boldsymbol{p})$:

$$\boldsymbol{f_{ij}}(\boldsymbol{p}) = \log\left(\begin{bmatrix} \frac{p_{i,j}}{1 - p_{i,j}} \\ \frac{p_{j,i}}{1 - p_{j,i}} \end{bmatrix}\right) \qquad A = I \bigotimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \qquad K = I \bigotimes \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

Using the probabilities of the model, it leads to:

$$\boldsymbol{f_{ij}}(\boldsymbol{p}) = \begin{bmatrix} -\log\theta + \log w_i - \log w_j \\ -\log\theta - \log w_i + \log w_j \end{bmatrix}$$

Davidson model has the following probabilities:

$$p_{i,j} = \frac{w_i}{w_i + w_j + \theta\sqrt{w_i w_j}} \qquad p_{i=j} = \frac{\theta\sqrt{w_i w_j}}{w_i + w_j + \theta\sqrt{w_i w_j}}$$

Let us define $\boldsymbol{f_{ij}}(\boldsymbol{p})$ and K such that $F(\boldsymbol{p}) = K\log(\boldsymbol{p})$:

$$\boldsymbol{f_{ij}}(\boldsymbol{p}) = \log\left(\begin{bmatrix} \frac{p_{i,j}}{p_{i=j}} \\ \frac{p_{j,i}}{p_{j=i}} \end{bmatrix}\right) \qquad K = I \bigotimes \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

Using the probabilities of the model, it leads to:

$$\boldsymbol{f_{ij}}(\boldsymbol{p}) = \begin{bmatrix} -\log\theta + \frac{1}{2}(\log w_i - \log w_j) \\ -\log\theta - \frac{1}{2}(\log w_i + \log w_j) \end{bmatrix}$$

For both ties models, the matrix X can be defined such that $F(\boldsymbol{p}) = X\begin{bmatrix} \log\theta \\ \boldsymbol{y} \end{bmatrix}$ with $y_i = \log w_i - \log w_N$ in the same way as the BTL model.

3.2.3 Mallows-Bradley-Terry triplet model

In the case of triple comparisons among N items, the number of categories, r, is 6 (for the number of possible rankings) and $n_s = \frac{N(N-1)(N-2)}{6}$, such that:

$$\boldsymbol{p'} = \begin{pmatrix} \boldsymbol{p'_{1,2,3}} & \boldsymbol{p'_{1,2,4}} & \cdots & \boldsymbol{p'_{2,3,4}} & \cdots & \boldsymbol{p'_{N-2,N-1,N}} \end{pmatrix}$$

$$\boldsymbol{p'_{i,j,k}} = \begin{pmatrix} p_{i,j,k} & p_{i,k,j} & p_{j,i,k} & p_{j,k,i} & p_{k,i,j} & p_{k,j,i} \end{pmatrix} \qquad \text{where} \qquad p_{i,j,k} = \frac{w_i^2 w_j}{D_{ijk}}$$

Let us define $\boldsymbol{f_{ijk}}(\boldsymbol{p})$ and B such that $F(\boldsymbol{p}) = K\log(\boldsymbol{p})$ with $K = I \otimes B$:

$$\boldsymbol{f_{ijk}}(\boldsymbol{p}) = \log\left(\begin{bmatrix} \frac{\sqrt{p_{i,j,k}}}{\sqrt{p_{k,j,i}}} \\ \frac{\sqrt{p_{i,k,j}}}{\sqrt{p_{j,k,i}}} \\ \frac{\sqrt{p_{j,i,k}}}{\sqrt{p_{k,i,j}}} \\ \frac{p_{i,j,k}}{p_{i,k,j}} \\ \frac{p_{j,i,k}}{p_{j,k,i}} \\ \frac{p_{k,i,j}}{p_{k,j,i}} \end{bmatrix}\right) \qquad B = \begin{bmatrix} \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{-1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{-1}{2} & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

The last row of B is optional to extract information symmetrically from first and second preferences, but B will be of rank 5 in any case. In the same way as the BTL model, the matrix X can be defined such that $F(\boldsymbol{p}) = X\boldsymbol{y}$ with $y_i = \log w_i - \log w_N$.

3.3 Eigenvector based method, rank centrality

This method, as described in [NOS12], aims to recover the stationary distribution of a Markov chain. A Markov chain is defined on a weighted directed graph. In the case of a BTL model, a graph can be described by using the N items as the nodes (states of

the Markov chain) with an edge between two items only if they are compared to each other.

The rank centrality method works by defining a random walk on this graph, which can be represented by a time-independent transition matrix P with $P_{ij} = \mathbb{P}[X_{t+1} = j | X_t = i]$ where X_t are the states of the Markov chain. This matrix can be chosen as follows:

$$P_{ij} = \begin{cases} \frac{h_{j,i}}{d_{\max}} & \text{if } i \neq j \\ 1 - \frac{1}{d_{\max}} \sum_{k \neq i} h_{k,i} & \text{if } i = j \end{cases}$$

where $d_{\max}$ is the maximum degree of a node in the simple graph and $h_{j,i}$ is the fraction of times the object j is chosen against the object i, i.e. $h_{j,i} = \frac{k_{j,i}}{k_{\boldsymbol{ij}}}$. This matrix P has the property that when the number of comparisons between two objects tends to infinity, the stationary distribution of the Markov chain is defined by $\pi_i = \frac{w_i}{\sum_j w_j}$.

In the finite case, we have that the stationary distribution of the random walk is the solution of

$$\hat{w}_i = \frac{\sum_{j \neq i} \hat{w}_j h_{j,i}}{\sum_{k \neq i} h_{k,i}}$$

3.4 Least Squares methods for BTL model

A more recent method, allowing us to recover the item weights based on observations, is to use a least squares method as described in [HOS20]. This method is similar to the one developed by [GSK69] but it has been proven that the more recent technique provides optimal results when the observations follow the BTL model.

As explained in [HOS20], let us define $R_{i,j} = \frac{h_{i,j}}{h_{j,i}}$ where $h_{i,j}$ (defined as in Chapter 3) is the fraction of times the object i is chosen against object j. $R_{i,j}$ converges to $\frac{w_i}{w_j}$ when the number of comparison between the two items is infinite since:

$$\lim_{k_{\boldsymbol{ij}} \to \infty} R_{i,j} = \lim_{k_{\boldsymbol{ij}} \to \infty} \frac{h_{i,j}}{h_{j,i}} = \frac{p_{i,j}}{p_{j,i}} = \frac{w_i}{w_j}$$

The quantities $R_{i,j}$ can be computed between all pairs of nodes. Since the number of comparisons is not infinite, those equations are approximately satisfied:

$$\log R_{i,j} \simeq \log w_i - \log w_j, \quad \text{for all pairs of items } (i,j)$$

A change of variables allows us to recover a linear system of equations in $\boldsymbol{z}$

$$\log R_{i,j} \simeq z_i - z_j, \quad \text{for all pairs of items } (i,j)$$

Zero values There is one problem that we need to take care of, namely the fact that $h_{a,b}$ might be equal to zero for some items a and b even if for a large value of $k_{\boldsymbol{ij}}$ this phenomena happens only with a small probability. When this happens, $h_{a,b}$ is given a small value, for example $\frac{1}{2k_{\boldsymbol{ab}}}$, and $h_{b,a}$ is modified (for example by subtracting $\frac{1}{2k_{\boldsymbol{ab}}}$) such that $h_{a,b} + h_{b,a} = 1$.

3.4.1 Least Squares method

Consider the simple graph $G = (V, E)$ where V represents the items and E is the set of edges linking two items that are compared to one another. The above equations can be formulated as in the following problem, where B' represents the incidence matrix of the graph G and $\boldsymbol{\epsilon}$ represents the error in the observations of $R_{i,j}$:

$$\log \boldsymbol{R} = B'\boldsymbol{z} + \boldsymbol{\epsilon}^1 \tag{3.1}$$

This problem (3.1) can be solved in the least squares sense:

$$\boldsymbol{z}^* = \arg \min_{z_1,\cdots,z_N} \sum_{(i,j)\in E} \Big(\log R_{i,j} - (z_i - z_j)\Big)^2$$

which is equivalent to:

$$BB'\boldsymbol{z}^* = B \log \boldsymbol{R}$$

An estimator $\hat{\boldsymbol{z}}$ of $\boldsymbol{z}$ is given by one solution of the above system:

$$\hat{\boldsymbol{z}} = \Big(BB'\Big)^\dagger B \log \boldsymbol{R}$$

where $A^\dagger$ denotes the Moore-Penrose pseudo-inverse of the matrix A. This solution is the one satisfying $\sum_{i=1}^N z_i = 0$. The weight given to each item can then be recovered using $\hat{w}_i = e^{\hat{z}_i}$.

3.4.2 Variance-Weighted Least Squares method

The least squares method does not allow us to consider the difference in the variance of the comparisons. Therefore, as it is suggested in [HOS20], we can opt for a weighted least squares method by dividing each term of the least squares problem by its variance. In that case, comparisons with a low variance will be given a higher weight in opposition to comparisons with a higher variance.

The problem remains to find the variances of $\log R_{i,j}$ for all pairs of items (i, j). By Taylor first order expansion of $\log h_1 - \log h_2 \approx \log p_1 - \log p_2 + (h_1 - p_1)\frac{1}{p_1} - (h_2 - p_2)\frac{1}{p_2}$, it holds that:

$$\mathbb{V}\left[\log \frac{h_{i,j}}{h_{j,i}}\right] \approx \frac{1}{p_{i,j}^2}\mathbb{V}\left[h_{i,j}\right] + \frac{1}{p_{j,i}^2}\mathbb{V}\left[h_{j,i}\right] - \frac{2}{p_{i,j}p_{j,i}}\mathbb{C}\left[h_{i,j}, h_{j,i}\right] \tag{3.2}$$

It holds that $k_* = k_{\boldsymbol{ij}} h_*$ follows a binomial distribution with a number of experiments equal to $k_{\boldsymbol{ij}}$ as h_* and k_* represent respectively the fraction of times and the number of times a ranking has been observed, meaning that

$$\mathbb{C}[h_{i,j}, h_{j,i}] = -\frac{1}{k_{\boldsymbol{ij}}} p_{i,j}p_{j,i} \text{ for } (i,j) \neq (j,i) \qquad \text{and} \qquad \mathbb{V}[h_{i,j}] = \frac{1}{k_{\boldsymbol{ij}}} p_{i,j}(1 - p_{i,j})$$

This leads to (calculation have been developed in [HOS20]):

$$\mathbb{V}(\log R_{i,j}) \approx V = \frac{1}{k_{\boldsymbol{ij}}} v_{ij} = \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{p_{j,i}}{p_{i,j}} + \frac{p_{i,j}}{p_{j,i}} + 2\right) = \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{1}{p_{i,j}} + \frac{1}{p_{j,i}}\right) \tag{3.3}$$

[1] $\log \boldsymbol{R}$ represents the element-wise logarithm of $\boldsymbol{R}$ containing the values $R_{i,j}$ for all the pairs of items i, j.

The least squares problem taking the variance into account is then given by:

$$\boldsymbol{z}^* = \arg\min_{z_1,\cdots,z_N} \sum_{(i,j)\in E} \frac{\Big(\log R_{i,j} - (z_i - z_j)\Big)^2}{\sqrt{v_{ij}/k_{\boldsymbol{ij}}}} \tag{3.4}$$

Solving the above problem is equivalent, in the least squares sense, to solve

$$V^{-1/2} \log \boldsymbol{R} = V^{-1/2} B' \boldsymbol{z}$$

where $V^{-1/2}$ is the diagonal matrix with the inverse of $\sqrt{v_{ij}/k_{\boldsymbol{ij}}}$ on the diagonal (with $V^{-1} = V^{-1/2}V^{-1/2}$), and from which one solution, chosen for an estimator of $\boldsymbol{z}$, is:

$$\hat{\boldsymbol{z}} = \log \hat{\boldsymbol{w}} = \left(BV^{-1}B'\right)^{\dagger} BV^{-1} \log \boldsymbol{R}$$

The weight given to each item can then again be recovered using $\hat{w}_i = e^{\hat{z}_i}$.

But in practice, the values of $w_i \ \forall i \in \{1, \cdots, N\}$ necessary to compute the probabilities $p_{i,j}$ in the variance's expression (3.3) are unknown. The method can thus not be used as it is, except with artificial weights to evaluate the method. An alternative could be to find an estimation of the variance, or iteratively find the variance and the weights.

Approximated variance The variance expression requires the values of $p_{i,j}$ that are unknown but for which we have an estimator, i.e. $h_{i,j}$, leading to an estimator of the variance equation (3.3):

$$\hat{V} = \frac{1}{k_{\boldsymbol{ij}}} \left(\frac{1}{h_{i,j}} + \frac{1}{h_{j,i}} \right)$$

This corresponds to the sample estimate of the variance as explained in [GSK69]:

$$\begin{aligned} \hat{V} &= \frac{\partial(\log R_{i,j})}{\partial \boldsymbol{h_{ij}}} \hat{V}(\boldsymbol{h_{ij}}) \left(\frac{\partial(\log R_{i,j})}{\partial \boldsymbol{h_{ij}}} \right)' \\ &= \frac{1}{h_{i,j}^2} \hat{V}\left(h_{i,j}\right) + \frac{1}{h_{j,i}^2} \hat{V}\left(h_{j,i}\right) - \frac{2}{h_{i,j}h_{j,i}} \hat{C}\left(h_{i,j}, h_{j,i}\right) \end{aligned}$$

where $\hat{V}(\boldsymbol{h_{ij}})$ is the sample variance of the vector $\boldsymbol{h_{ij}} = \begin{bmatrix} h_{i,j} & h_{j,i} \end{bmatrix}$ such that

$$\hat{C}(h_{i,j}, h_{j,i}) = -\frac{1}{k_{\boldsymbol{ij}}} h_{i,j}h_{j,i} \text{ for } (i,j) \neq (j,i) \qquad \text{and} \qquad \hat{V}(h_{i,j}) = \frac{1}{k_{\boldsymbol{ij}}} h_{i,j}(1 - h_{i,j})$$

This is indeed equivalent to equation (3.2) by replacing $p_{i,j}$ by $h_{i,j}$ and the variance matrix V by the sample variance matrix $\hat{V}$ of the vector.

Iterative method To estimate the variance of $\log \boldsymbol{R}$, we first set $v_{ij} = 1$ which allows us to solve (3.4) and find values for $\boldsymbol{w}$. We can then improve the values of v_{ij} using $\boldsymbol{w}$ by computing the values of $p_{i,j}$. This process is done several times until convergence.

Chapter 4
Least Squares methods for BTL extensions models

As they did in [HOS20] for the BTL model, we can develop a least squares method to estimate the item scores for extensions of the BTL models, in particular for pairwise comparisons with ties using Rao&Kupper and Davidson models and for triple comparisons using Plackett-Luce and Mallows-Bradley-Terry models.

As we will see in the next Chapters, the solutions to the weighted least squares method are optimal for the considered problems. This Chapter aims at stating the equations one can use in order to recover the weights based on observations of item comparisons as an extension to the method developed for the BTL model.

4.1 General methods

For the BTL model, $R_{i,j}$ has been defined as $\frac{h_{i,j}}{h_{j,i}}$ where $h_{i,j}$ is the fraction of times the object i is chosen against object j or also fraction of times ranking i, j is chosen against ranking j, i, and the approximately satisfied equation for each pair i, j was

$$\log\left(\frac{h_{i,j}}{h_{j,i}}\right) \approx \log\left(\frac{p_{i,j}}{p_{j,i}}\right) = \log(w_i) - \log(w_j) \tag{4.1}$$

This idea to have equations linking item scores to fraction of times a certain choice is made can be used to extend the least squares method for the other models presented in Part I. The models used in this Chapter could have additional parameters, let those parameters be gathered in the vector $\boldsymbol{\theta}$.

The general aim is thus to find new equations approximately satisfied for the different models such that we also end up with a system of linear equations in the logarithm of the parameters $\log \boldsymbol{\theta}$ and $\log \boldsymbol{w}$ (the item weights) keeping as left term, a ratio of the form $\frac{h_\alpha}{h_\beta}$ where h_α and h_β represent the fraction of times an event occurs and p_α and p_β is the probability that this event occurs. We focus thus the extended method on general equations of the type

$$\lim_{k\to\infty} \frac{h_\alpha}{h_\beta} = \frac{p_\alpha}{p_\beta} = \prod_{b=1}^{\tilde{N}-N} (\theta_b)^{\xi_b} \prod_{a=1}^{N} (w_a)^{\zeta_a}$$

where $\tilde{N}$ is the total number of parameters, N is the number of items and thus $\tilde{N} - N$ is the number of additional parameters.

Since the number of comparisons is finite, the left term is approximately equal to the right term and by taking the logarithm of the expression, it becomes then linear in $\log \boldsymbol{\theta}$ and $\log \boldsymbol{w}$:

$$\log \frac{h_\alpha}{h_\beta} \approx \sum_{b=1}^{\tilde{N}-N} \xi_b \log \theta_b + \sum_{a=1}^{N} \zeta_a \log w_a \tag{4.2}$$

4.1.1 Least Squares method

The same change of variable as for the BTL model, i.e., $\boldsymbol{z} = \begin{bmatrix}\log(\boldsymbol{\theta}) & \log(\boldsymbol{w})\end{bmatrix}'$ where $\boldsymbol{w}$ represents the item weights vector and $\boldsymbol{\theta}$ any additional parameters, allows us to recover a linear system of equations in $\boldsymbol{z}$ from all the equations of (4.2):

$$\log \boldsymbol{R} = \boldsymbol{Y} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ represents the error in the functions of observations h_* and M' is the matrix of coefficients for the corresponding equations depending on the model. Similarly to BTL system, $\boldsymbol{R}$ is the vector of ratio of h_*.

The idea of the method is to solve the above system in the least squares sense, which is equivalent to find $\boldsymbol{z}^*$ such that:

$$MM'\boldsymbol{z}^* = M\boldsymbol{Y}$$

A least squares estimator $\hat{\boldsymbol{z}}$ of $\boldsymbol{z}$ is given by one of the solutions of the above system:

$$\hat{\boldsymbol{z}} = \left(MM'\right)^\dagger M\boldsymbol{Y}$$

The weight given to each item can then be recovered using $\hat{w}_i = e^{\log \hat{w}_i}$ with $\hat{\boldsymbol{z}} = \begin{bmatrix}(\log \hat{\boldsymbol{\theta}})' & (\log \hat{\boldsymbol{w}})'\end{bmatrix}'$.

Some h_π, for a ranking π, could be equal to zero. As for the BTL model in Section 3.4, those values will be replaced according to the model used.

4.1.2 Covariance-Weighted Least Squares method

As for the BTL least squares method, the previous method does not allow us to consider the difference in the variance of the comparisons. Indeed, for the BTL method, they divided each equation by its variance to get a weighted least squares problem. As it is done for BTL, we can also opt for a weighted least squares method by dividing each term of the least squares problem by its variance. For the BTL least squares method, dividing by the variance was equivalent to multiply by the inverse of the covariance matrix because this matrix was diagonal. Multiplying by the inverse of the covariance matrix is thus another option to consider.

To recover the item weight, the following system of equations will be solved using the weighted least squares method:

$$(W^\dagger)^{1/2} \log \boldsymbol{R} \approx (W^\dagger)^{1/2} M'\boldsymbol{z}^*$$

where the weight matrix W such that $W^\dagger = (W^\dagger)^{-1/2}(W^\dagger)^{-1/2}$ equals to:

- The covariance matrix of $\log \boldsymbol{R}$, or
- The diagonal matrix with the variance vector of $\log \boldsymbol{R}$ on the diagonal

The least squares estimator $\hat{\boldsymbol{z}}$ of $\boldsymbol{z}$ is then given by one solution of the system $MW^\dagger \log \boldsymbol{R} = MW^\dagger M' \boldsymbol{z}^*$:

$$\hat{\boldsymbol{z}} = \left(MW^\dagger M'\right)^\dagger MW^\dagger \log \boldsymbol{R}$$

The weight given to each item can then again be recovered using $\hat{w}_i = e^{\log \hat{w}_i}$ with $\hat{\boldsymbol{z}} = \left[(\log \hat{\boldsymbol{\theta}})' \quad (\log \hat{\boldsymbol{w}})'\right]'$. Results about the performances of that solution minimizing $(W^\dagger)^{1/2} \log \boldsymbol{R} - (W^\dagger)^{1/2} M' \boldsymbol{z}^*$ in the least squares sense will be provided in Chapter 6.

Approximated variance $\boldsymbol{V}$ The variance of $\boldsymbol{Y}$ is unknown but can be estimated using a Taylor expansion of the expression as it was done for the BTL method. The variance is a block diagonal matrix, as two rankings on different items will have independent h. Remember that the equations will always have a ratio of h as left term. By Taylor first order expansion of $\log\left(\frac{h_1}{h_2}\right) = \log h_1 - \log h_2 \approx \log p_1 - \log p_2 + (h_1 - p_1)\frac{1}{p_1} - (h_2 - p_2)\frac{1}{p_2}$ and using covariance properties, it holds that:

$$\begin{aligned}\mathbb{C}\left[\log \frac{h_1}{h_2}, \log \frac{h_3}{h_4}\right] &\approx \mathbb{C}\left[\frac{h_1}{p_1} - \frac{h_2}{p_2}, \frac{h_3}{p_3} - \frac{h_4}{p_4}\right] \\ &= \frac{1}{p_1 p_3}\mathbb{C}\left[h_1, h_3\right] - \frac{1}{p_1 p_4}\mathbb{C}\left[h_1, h_4\right] - \frac{1}{p_2 p_3}\mathbb{C}\left[h_2, h_3\right] + \frac{1}{p_2 p_4}\mathbb{C}\left[h_2, h_4\right]\end{aligned} \tag{4.3}$$

This expression will be further developed regarding the model used. Note that in the above equation, h_a represents the fraction of times a certain event a occurs. It could be that in h_a, a represents a certain ranking, but one could also have that h_a is the sum of the fraction of times different rankings occur, such as $h_{i,.,.} = h_{i,j,k} + h_{i,k,j}$ for example in triple comparisons.

Sample approximated variance $\hat{\boldsymbol{V}}$ The covariance expressions require the values of the different probabilities of observing the rankings, $\boldsymbol{p_{ij}}$ or $\boldsymbol{p_{ijk}}$, that are unknown in reality. We have however an estimator for those probabilities, i.e., $\boldsymbol{h_{ij}}$ or $\boldsymbol{h_{ijk}}$, leading to an estimator of the covariance equations by replacing p_* by h_* in the equations. This corresponds to the sample estimate of the covariance:

$$\hat{V} = \frac{\partial(\log \boldsymbol{R})}{\partial \boldsymbol{h}} \hat{V}(\boldsymbol{h}) \left(\frac{\partial(\log \boldsymbol{R})}{\partial \boldsymbol{h}}\right)'$$

where $\hat{V}(\boldsymbol{h})$ is the block diagonal matrix with the blocks $\hat{V}(\boldsymbol{h_{ijk}}/\boldsymbol{h_{ij}})$, the sample variances of the vectors $\boldsymbol{h_{ijk}} = \begin{bmatrix} h_{i,j,k} & h_{i,k,j} & h_{j,i,k} & h_{j,k,i} & h_{k,i,j} & h_{k,j,i} \end{bmatrix}$, $\boldsymbol{h_{ij}} = \begin{bmatrix} h_{i,j} & h_{j,i} & h_{i=j} \end{bmatrix}$.

The approximation of the variance and its estimation has been compared with the empirical variance on Figures A.1 and A.2 in Appendix A for the four models considered.

Iterative method As it was done for the BTL model, we can estimate the covariance matrix using an iterative method. To do so, let us take the identity matrix as first guess of the covariance. With the matrix being fixed, the system of equations can be solved, and we obtain a vector $\boldsymbol{w}$. With this vector, it is possible to compute the different probabilities (as explained and detailed in Part I) and we obtain thus another covariance matrix. The process can be repeated until convergence.

4.2 Application to BTL models with ties

The aim of this Section is to develop a method able to recover item weights based on pairwise comparisons in the case where ties are allowed. In Part I, two models were presented, namely the Rao & Kupper and the Davidson models. The method developed will be based on the one of Section 3.4. Remember that in the case where ties are allowed, a factor θ is added to the model. In this optic, we are looking for linear functions in the logarithm of the parameters $\log \boldsymbol{w}$ and $\log \theta$.

Equation (4.1) for each pair i, j can be extended for pairwise comparison with ties in two different ways for each pair i, j:

- Ratio of fractions of times a choice is done against another:

$$\frac{h_{i,j}}{h_{j,i}} \qquad \frac{h_{i,j}}{h_{i=j}} \qquad \frac{h_{j,i}}{h_{i=j}} \tag{4.4}$$

- Ratio of factions of times an item i is preferred against when it is not strictly preferred:

$$\frac{h_{i,j}}{1-h_{i,j}} \qquad \frac{h_{j,i}}{1-h_{j,i}} \tag{4.5}$$

Since $h_{i,j}, h_{i=j}$ tend to $p_{i,j}, p_{i=j}$ for a large number of comparisons, we can approximate the previous equations by functions of the item scores w_i, w_j and ties parameter θ for the two models presented in Chapter 2. We will focus on the functions which are linear regarding the element-wise logarithm of the parameters since some equations also contain logarithmic terms of linear or quadratic functions of the model parameters.

Zero values We need to take care of the same problem as for the BTL equations of Chapter 3.4, namely the fact that some h_π from $h_{i,j}, h_{j,i}, h_{i=j}$ might be equal to zero for some items i, j even for a large value of $k_{\boldsymbol{ij}}$. When this happens, h_π is given a small value, for example $\frac{1}{3k_{\boldsymbol{ij}}}$, and the non-zero ones are modified such that they sum to one.

BTL-transformed method With pairwise comparison information with ties, one could also recover pairwise comparison information with which the BTL methods can be used.

It indeed holds that if the values of $h_{i,j}, h_{j,i}, h_{i=j}$ are known, we can define pairwise comparison information in two different ways either by splitting the ties observations

between the two items or by ignoring the ties observations:

$$\tilde{h}_{i,j} = h_{i,j} + \frac{h_{i=j}}{2} \qquad \tilde{h}_{i,j} = \frac{h_{i,j}}{h_{i,j} + h_{j,i}}$$
$$\tilde{h}_{j,i} = h_{j,i} + \frac{h_{i=j}}{2} \qquad \tilde{h}_{j,i} = \frac{h_{j,i}}{h_{i,j} + h_{j,i}}$$
$$\tilde{k}_{\boldsymbol{ij}} = k_{\boldsymbol{ij}} \qquad \tilde{k}_{\boldsymbol{ij}} = (h_{i,j} + h_{j,i})k_{\boldsymbol{ij}}$$

From which we can apply the (weighted) least squares approaches for pairwise comparison and the zero values corrections can be applied on $\tilde{h}_{i,j}$ instead of $h_{i,j}$.

4.2.1 Rao & Kupper

Remember that the model is given by

$$p_{i,j} = \frac{w_i}{w_i + \theta w_j} \qquad p_{i=j} = \frac{(\theta^2 - 1)w_i w_j}{(w_i + \theta w_j)(w_j + \theta w_i)}$$

By observing that $\frac{p_{i,j}}{1-p_{i,j}} = \frac{w_i}{\theta w_j}$, we end up with two equations per pair of items i, j:

$$\log \frac{h_{i,j}}{1 - h_{i,j}} \approx \log w_i - \log w_j - \log \theta$$
$$\log \frac{h_{j,i}}{1 - h_{j,i}} \approx \log w_j - \log w_i - \log \theta$$

To be able to develop a weighted least squares method as explained before, we need to compute the variance matrix of the left member $\boldsymbol{Y}$ of the system of equations described above. Using the relation (4.3) we obtain

$$\mathbb{C}\left[\log \frac{h_{i,j}}{1 - h_{i,j}}, \frac{h_{j,i}}{1 - h_{j,i}}\right] \approx \frac{\mathbb{C}\left[h_{i,j}, h_{j,i}\right]}{p_{i,j}p_{j,i}} - \frac{\mathbb{C}\left[h_{i,j}, 1 - h_{j,i}\right]}{p_{i,j}(1 - p_{j,i})} - \frac{\mathbb{C}\left[1 - h_{i,j}, h_{j,i}\right]}{(1 - p_{i,j})p_{j,i}} + \frac{\mathbb{C}\left[1 - h_{i,j}, 1 - h_{j,i}\right]}{(1 - p_{i,j})(1 - p_{j,i})}$$

$$\mathbb{V}\left[\log \frac{h_{i,j}}{1 - h_{i,j}}\right] \approx \frac{\mathbb{C}\left[h_{i,j}, h_{i,j}\right]}{p_{i,j}p_{i,j}} - 2\frac{\mathbb{C}\left[h_{i,j}, 1 - h_{i,j}\right]}{p_{i,j}(1 - p_{i,j})} + \frac{\mathbb{C}\left[1 - h_{i,j}, 1 - h_{i,j}\right]}{(1 - p_{i,j})(1 - p_{i,j})}$$

As $k_{i,j} = k_{\boldsymbol{ij}}h_{i,j}$ (resp. $k_{j,i} = k_{\boldsymbol{ij}}h_{j,i}$) follows a multinomial distribution[1] with a number of experiments equal to $k_{\boldsymbol{ij}}$, it holds that

$$\mathbb{C}[h_{i,j}, h_{j,i}] = -\frac{1}{k_{\boldsymbol{ij}}}p_{i,j}p_{j,i} \text{ for } (i,j) \neq (j,i) \qquad \text{and} \qquad \mathbb{V}[h_{i,j}] = \frac{1}{k_{\boldsymbol{ij}}}p_{i,j}(1 - p_{i,j})$$

The above expressions become thus

$$\mathbb{C}\left[\log \frac{h_{i,j}}{1 - h_{i,j}}, \frac{h_{j,i}}{1 - h_{j,i}}\right] \approx \frac{-1}{k_{\boldsymbol{ij}}} \frac{1}{(1 - p_{i,j})(1 - p_{j,i})}$$
$$\mathbb{V}\left[\log \frac{h_{i,j}}{1 - h_{i,j}}\right] \approx \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{1}{1 - p_{i,j}} + \frac{1}{p_{i,j}}\right)$$

[1]The multinomial distribution is the generalization of the binomial distribution for a set of r mutually exclusive events. In other words, $k_{i,j} \sim \mathcal{M}(k_{\boldsymbol{ij}}, r, p_1, \cdots, p_k)$ where p_i is the probability that an event i occurs, and $\mathcal{B}(m,p) = \mathcal{M}(m, 2, p, 1-p)$.

4.2.2 Davidson

This model is defined by

$$p_{i,j} = \frac{w_i}{w_i + w_j + \theta\sqrt{w_i w_j}} \qquad p_{i=j} = \frac{\theta\sqrt{w_i w_j}}{w_i + w_j + \theta\sqrt{w_i w_j}}$$

The same reasoning as in the BTL model can be used for the Davidson model. Three equations are found for one pair of items. Those relations are given by

$$\begin{aligned}
\log\frac{h_{i,j}}{h_{j,i}} &\approx \log w_i - \log w_j \\
\log\frac{h_{i=j}}{h_{i,j}} &\approx \log\theta + \frac{1}{2}\log w_j - \frac{1}{2}\log w_i \\
\log\frac{h_{i=j}}{h_{j,i}} &\approx \log\theta + \frac{1}{2}\log w_i - \frac{1}{2}\log w_j
\end{aligned}$$

Developing the expression (4.3) for the different cases, knowing that $k_{i,j}, k_{j,i}$ and $k_{i=j}$ follow multinomial distribution, allow us to obtain

$$\begin{aligned}
\mathbb{C}\left[\log\frac{h_{i,j}}{h_{j,i}}, \log\frac{h_{i=j}}{h_{i,j}}\right] &\approx \frac{-1}{k_{\boldsymbol{ij}}}\frac{1}{p_{i,j}} & \mathbb{V}\left[\log\frac{h_{i,j}}{h_{j,i}}\right] &\approx \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{1}{p_{i,j}} + \frac{1}{p_{j,i}}\right) \\
\mathbb{C}\left[\log\frac{h_{i,j}}{h_{j,i}}, \log\frac{h_{i=j}}{h_{j,i}}\right] &\approx \frac{1}{k_{\boldsymbol{ij}}}\frac{1}{p_{j,i}} & \mathbb{V}\left[\log\frac{h_{i=j}}{h_{i,j}}\right] &\approx \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{1}{p_{i=j}} + \frac{1}{p_{i,j}}\right) \\
\mathbb{C}\left[\log\frac{h_{i=j}}{h_{i,j}}, \log\frac{h_{i=j}}{h_{j,i}}\right] &\approx \frac{1}{k_{\boldsymbol{ij}}}\frac{1}{p_{i=j}} & \mathbb{V}\left[\log\frac{h_{i=j}}{h_{j,i}}\right] &\approx \frac{1}{k_{\boldsymbol{ij}}}\left(\frac{1}{p_{i=j}} + \frac{1}{p_{j,i}}\right)
\end{aligned}$$

4.3 Application to triple comparison models

We will consider observations/data of ranking of three items. For each triplet of objects i, j, k, observations consist of the fraction of times each possible ranking has been chosen, i.e. $h_{i,j,k}, h_{i,k,j}, h_{j,i,k}, h_{j,k,i}, h_{k,i,j}, h_{k,j,i}$. We might also consider other types of observations/data for three items: choosing the favorite item among the three or choosing the two favorite items among the three (or equivalently selecting the less preferred item).

Equation (4.1) for each pair i, j can be extended for triple comparisons in three different ways for each triplet i, j, k:

- Ratio of fractions of times an object i is ranked before j (resp. i before k or j before k) is chosen against j before i (resp. k before i or k before j):

$$\log\left(\frac{h_{i,j}}{h_{j,i}}\right) = \log\left(\frac{h_{i,j,k} + h_{i,k,j} + h_{k,i,j}}{h_{j,i,k} + h_{j,k,i} + h_{k,j,i}}\right) \tag{4.6}$$

- Ratio of fractions of times an object i has a certain rank against an object j (resp. i against k or j against k):
  - First rank

$$\log\left(\frac{h_{i,\cdot,\cdot}}{h_{j,\cdot,\cdot}}\right) = \log\left(\frac{h_{i,j,k} + h_{i,k,j}}{h_{j,i,k} + h_{j,k,i}}\right) \tag{4.7}$$

– Second rank

$$\log\left(\frac{h_{\cdot,i,\cdot}}{h_{\cdot,j,\cdot}}\right) = \log\left(\frac{h_{j,i,k} + h_{k,i,j}}{h_{i,j,k} + h_{k,j,i}}\right) \tag{4.8}$$

– Last rank

$$\log\left(\frac{h_{\cdot,\cdot,i}}{h_{\cdot,\cdot,j}}\right) = \log\left(\frac{h_{j,k,i} + h_{k,j,i}}{h_{i,k,j} + h_{k,i,j}}\right) \tag{4.9}$$

- Ratio of fractions of times a ranking is chosen against another:

$$\log\left(\frac{h_\pi}{h_{\pi'}}\right) \tag{4.10}$$

Since $h_{i,j,k}$ tends to $p_{i,j,k}$ for a large number of comparisons, we can approximate the previous equations by functions of the item scores w_i, w_j, w_k for the two models for triple comparisons. We will focus on the functions which are linear regarding the logarithm parameters $\log \boldsymbol{w}$ since some equations also contain logarithmic terms of linear or quadratic functions of the three parameters w_i, w_j, w_k. The equations giving the best empirical performances (shown in Figure A.3 in Appendix A) will be kept.

Zero values We need to take care of the same problem as for the BTL equations of Chapter 3.4, namely the fact that some h_π from $h_{i,j,k}, h_{i,k,j}, h_{j,i,k}, h_{j,k,i},\ h_{k,i,j}, h_{k,j,i}$ might be equal to zero for some items i, j and k, even for a large value of $k_{\boldsymbol{ijk}}$. When this happens, h_π is given a small value, for example $\frac{1}{6k_{\boldsymbol{ijk}}}$, and the non-zero ones are modified such that they sum to one.

BTL-transformed method With triple comparison information, one can also recover pairwise comparison information with which the BTL methods can be used.

It indeed holds that if the values of $h_{i,j,k}, h_{i,k,j}, h_{j,i,k}, h_{j,k,i}, h_{k,i,j}, h_{k,j,i}$, are known, we can define:

$$\begin{aligned} h^k_{i,j} &= h_{i,j,k} + h_{i,k,j} + h_{k,i,j} & \qquad h^k_{j,i} &= h_{j,i,k} + h_{j,k,i} + h_{k,j,i} \\ h^j_{i,k} &= h_{i,j,k} + h_{i,k,j} + h_{j,i,k} & \qquad h^j_{k,i} &= h_{j,k,i} + h_{k,i,j} + h_{k,j,i} \\ h^i_{j,k} &= h_{j,i,k} + h_{j,k,i} + h_{i,j,k} & \qquad h^i_{k,j} &= h_{i,k,j} + h_{k,i,j} + h_{k,j,i} \end{aligned}$$

with $h^k_{i,j}$ being the fraction of times the object i has been ranked before object j in the triplet ijk.

And it results that:

$$h_{i,j} = \frac{1}{k_{\boldsymbol{ij}}} \sum_{k \neq i,j} k_{\boldsymbol{ijk}} h^k_{i,j} \quad \text{with } k_{\boldsymbol{ij}} = \sum_{k \neq i,j} k_{\boldsymbol{ijk}}$$

From which we can apply the (weighted) least squares approaches for pairwise comparison and the zero value corrections can be applied on $h_{i,j}$ instead of $h_{i,j,k}$.

4.3.1 Mallow-Bradley-Terry

Regarding this model, only the equations from (4.10) are linear in $\log w_i, \log w_j, \log w_k$:

$$\log\left(\frac{h_\pi}{h_{\pi'}}\right) \approx (\pi'^{-1} - \pi^{-1})(i)\log(w_i) + (\pi'^{-1} - \pi^{-1})(j)\log(w_j) + (\pi'^{-1} - \pi^{-1})(k)\log(w_k)$$
$$\text{for all rankings } \pi, \pi' \in \{(i,j,k),(i,k,j),(j,i,k),(j,k,i),(k,i,j),(k,j,i)\}$$
$$\text{for all triplet of items } (i,j,k) \tag{4.11}$$

where $\pi^{-1}(i)$ is the rank of the object i in the ranking π. For example,

$$\log\left(\frac{h_{i,j,k}}{h_{k,i,j}}\right) \approx \log\left(\frac{p_{i,j,k}}{p_{k,i,j}}\right) = \log\left(\frac{\frac{w_i^2 w_j}{D_{ijk}}}{\frac{w_k^2 w_i}{D_{ijk}}}\right) = \log\left(\frac{w_i^2}{w_i}\right) + \log(w_j) + \log\left(\frac{1}{w_k^2}\right)$$
$$= 1\log(w_i) + 1\log(w_j) - 2\log(w_k)$$

This gives 15 equations [2] for each triplet of items including 5 linearly independent (since there are 6 values of h_π and they sum to one).

Let us now compute the variance of $\log(h_\pi/h_{\pi'})$ for some rankings π and π'. Let α, β, γ and δ be some rankings of the items i, j, k. We are then looking for $\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta}, \log\frac{h_\gamma}{h_\lambda}\right]$. It holds that $k_\alpha = k_{\boldsymbol{ijk}} h_\alpha$ follows a multinomial distribution with a number of experiments equal to $k_{\boldsymbol{ijk}}$ as h_α represents the fraction of times and k_α the number of times the ranking α has been observed. This means that

$$\mathbb{C}[h_\alpha, h_\beta] = -\frac{1}{k_{\boldsymbol{ijk}}} p_\alpha p_\beta \text{ for } \alpha \neq \beta \qquad \text{and} \qquad \mathbb{V}[h_\alpha] = \frac{1}{k_{\boldsymbol{ijk}}} p_\alpha (1 - p_\alpha)$$

The computation of (4.3) occurs to be similar in several possibilities depending on if some rankings α, β, γ and δ are the same ones[3]. We will develop the cases where the four triplets are different ($\alpha \neq \beta \neq \gamma \neq \lambda$), the two numerators are equal ($\alpha = \gamma,\ \alpha \neq \beta \neq \lambda$), the two denominators are equal ($\beta = \lambda,\ \beta \neq \alpha \neq \gamma$), one numerator equals the denominator of the other term ($\alpha = \lambda,\ \alpha \neq \beta \neq \gamma$) and the two ratios are equal ($\alpha = \gamma$ and $\beta = \lambda$ for $\alpha \neq \beta$). Details of the computations can be found in Appendix B, and we obtain:

[2] In reality, there are 30 combinations of two rankings but $\log\left(\frac{h_\pi}{h_{\pi'}}\right) = -\log\left(\frac{h_{\pi'}}{h_\pi}\right)$ thus we only need the 15 pairs of two rankings.

[3] In the equations used, the left member will never be equal to 1 meaning that it always holds that $\alpha \neq \beta$ and $\gamma \neq \lambda$.

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta}, \log\frac{h_\gamma}{h_\lambda}\right] \approx \frac{1}{k_{\boldsymbol{ijk}}} \begin{cases} \left(\frac{1}{p_\alpha} + \frac{1}{p_\beta}\right) & \text{if } \alpha = \gamma, \beta = \lambda \\ \frac{1}{p_\alpha} & \text{if } \alpha = \gamma \\ \frac{1}{p_\beta} & \text{if } \beta = \lambda \\ -\frac{1}{p_\alpha} & \text{if } \alpha = \lambda \\ 0 & \text{otherwise} \end{cases} \tag{4.12}$$

4.3.2 Plackett-Luce

For this model, equations (4.6) and (4.7) are linear in $\log w_i$ (some equations from (4.10) are linear but not for all ranking pairs).

$$\log\left(\frac{h_{i,j}}{h_{j,i}}\right) \approx \log(w_i) - \log(w_j) \tag{4.13}$$
$$\log\left(\frac{h_{i,\cdot,\cdot}}{h_{j,\cdot,\cdot}}\right) \approx \log(w_i) - \log(w_j)$$
$$\text{for all pairs } (i,j) \in \{(i,j),(i,k),(j,k)\}, \text{ for all triplet of items } (i,j,k)$$

This gives, for both, 3 equations for each triplet of items including 3 linearly independent for the first one and 2 for the second. The set of equations kept for the analysis is the first one. It is indeed the one that is the most informative among the available data as $h_{i,j}$ hides a sum of three h_* with $*$ being a ranking of three items i, j, k. The corrections applied on the observations when $h_* = 0$ will indeed be more attenuated thanks to the sum while in the other sets of equations, only a portion of information is kept, and the equations are therefore not as accurate as the ones from the first set of equations. The effect of the modifications on the observations will have an impact more important.

With the system of equations known, one can recover the variance matrix of $\boldsymbol{Y}$. Let α and β represent the relative order of two items of the triplet i, j, k. We could for example take $\alpha = i, j$ meaning that item i is ranked above item j. Denote by α^{-1} and β^{-1} the opposite relative order of the two items. Taking the same example for α, we would have $\alpha^{-1} = j, i$.

With the definition of α and β, we do not have that h_α follows a multinomial distribution, as it was the case for the variance of $\boldsymbol{Y}$ in the equations of the Mallows-Bradley-Terry model. To go further in the computations, it is thus necessary to consider all the different combinations of α and β one can observe. Details of the computations

can be found in Appendix B, and we obtain:

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_{\alpha^{-1}}},\log\frac{h_\beta}{h_{\beta^{-1}}}\right]\approx\frac{1}{k_{\boldsymbol{ijk}}}\begin{cases}\dfrac{1}{p_{i,j}p_{i,k}}\left(\dfrac{p_{j,k,i}+p_{k,j,i}}{p_{k,i}p_{j,i}}-1\right) & \text{if } \alpha=i,j \text{ and } \beta=i,k\\ \dfrac{1}{p_{i,j}p_{j,k}}\left(\dfrac{p_{k,j,i}}{p_{k,j}p_{j,i}}-1\right) & \text{if } \alpha=i,j \text{ and } \beta=j,k\\ \dfrac{1}{p_{i,k}p_{j,k}}\left(\dfrac{p_{k,i,j}+p_{k,j,i}}{p_{k,j}p_{k,i}}-1\right) & \text{if } \alpha=i,k \text{ and } \beta=j,k\\ \left(\dfrac{1}{p_\alpha}+\dfrac{1}{p_{\alpha^{-1}}}\right) & \text{if } \alpha=\beta\end{cases} \tag{4.14}$$

Part III

Best Linear Unbiased Estimators in Comparison Models

The problem of finding parameters estimators based on noisy observations linearly dependent of those parameters is a more general problem which has been widely studied. This corresponds to approximate an infeasible linear system of equations, and we will focus on the basic method of solving it in the least squares sense.

Approximating this system allows, under certain conditions, to find the best linear unbiased estimator (BLUE) of the unknown. This means that the estimator has the smallest variance among the ones that are unbiased and linear in the output variables. This is the theorem stated in Section 5.1 derived by Gauss and Markov. The problem is to estimate β such that $Y = X'\beta + \epsilon$ with a known full rank matrix X', observations Y and an unknown error vector ϵ which is homoscedastic (the entry of the error vector all have the same finite variance) and uncorrelated (covariance between two different entries of the error vector is null) and the estimator is given by $\hat{\beta} = (XX')^{-1} XY$. Note that this theorem assumes that the noise ϵ has a known and well-defined variance matrix. If those conditions are not satisfied, the statement of the problem will need some modifications if one hopes to find a solution to the problem.

The above problem with a full rank matrix X' is well known. However, in the case of a non-full rank matrix X', the problem becomes more difficult as the parameter β becomes unobservable and there is no linear unbiased estimator of the parameters. Take, for example, a problem with four recipients of a liquid with different concentrations c_i of a solute. If we observe the difference in concentration between each solution, there does not exist a unique approximation of their concentration. The system of equations of this problem consists indeed in 6 equations, where the matrix of coefficients is not of full rank.

In mathematical words, unobservable parameters means that there exists no matrix B' satisfying $\mathbb{E}[\hat{\beta}] = \mathbb{E}[B'Y] = \mathbb{E}[B'(X'\beta + \epsilon)] = B'X'\mathbb{E}[\beta] = \beta$. There is in fact a class of β satisfying the equations $Y = X'\beta + \epsilon$ which is thus a class of equivalence for this problem. In our example, each concentration is defined up to an additive factor, and there is thus a class of concentrations satisfying the problem.

Two different approaches will be developed in Section 5.2 to propose an alternative to the initial problem.

- The first one ensues from [GZ64] and consists of defining a smaller number of new observable parameters, functions of the initial parameters. A variable γ is introduced as linear combinations of the parameter β such that $\gamma = L'\beta$ is observable. It will be shown that in this case, L must belong to the columnspace of X. The set of solutions to this problem is given by $\hat{\gamma} = L'\hat{\beta}$ where $\hat{\beta} = (XX')^{-}XY + \left(I - (XX')^{-}XX'\right)z$ for all vectors z where A^{-} is a generalized inverse of A satisfying (a) $AA^{-}A = A$, (b) $A^{-}AA^{-} = A^{-}$ and (c) $(AA^{-})' = AA^{-}$. For our previous example, the new observable parameters could be the estimators of the differences in concentration between solutions $\gamma_1 = c_1 - c_2$, $\gamma_2 = c_2 - c_3$ and $\gamma_3 = c_3 - c_4$.
- Our contribution consists in another approach which permits to have a solvable problem. To do so we add mandatory conditions on the parameters such as $\beta^p \in C = \{x | S'x = m\}$ for some matrix S such that $\begin{bmatrix} X & S \end{bmatrix}$ has full rank. It means that we choose a representative of the class of equivalence defined by $Y = X'\beta + \epsilon$. In our example, we can consider the concentrations of the solutions for a concentration $c_1 = 0$ mg/L for the first solution. In this case and when the error vector is as before, the estimator is chosen among the vectors that minimize the sum of the squares of the residuals, i.e. $XX'\hat{\beta} = XY$. The BLUE estimator of β^p is given by $\hat{\beta}_c = (Q'QXX'Q'Q)^{\dagger}XY + B(S'B)^{-1}m$ where Q' is an orthonormal basis of $\mathrm{Ker}(S')$ and B a basis of $\mathrm{Ker}(X')$ as shown in Section 5.2.2.

When the problem is such that the observations Y are non-independent and/or the errors are non-homoscedastic, the error vector will not satisfy the conditions of the Gauss-Markov theorem anymore. In a similar manner as in [GZ64], we have showed in Section 5.3 that the system can be reduced to a new system $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ respecting the hypothesis of the problems solved in Section 5.2. All the results presented in the first Sections can thus be used for solving this new approximation problem.

When we redefine the problem in order to have a unique vector of true weights β^p, we do not make the estimator of that vector unique. There exists indeed an infinite number of constraints satisfied by that vector. Take our previous example with the true concentrations given by $\beta^p = \begin{bmatrix} 0 & 1 & 2 & 3 \end{bmatrix}$, one could have added the constraint $\beta^p_1 = 0$ mg/L or $\sum \beta^p = 6$ mg/L to the initial problem. Regarding the added constraint, the BLUE estimator of β^p is different. By imposing that the estimator belongs to a constrained subspace $C = \{x | S'x = m\}$, the estimator of β^p is render unique, and we have shown in Section 5.3.1 that the estimator $\hat{\beta}^p_c$ minimizes in that case a particular weighted 2−norm defined by $\left\|\beta^p - \hat{\beta}\right\|_{2,A} = \sqrt{(\beta^p - \hat{\beta})'A(\beta^p - \hat{\beta})}$ with A a positive

definite matrix for all $\hat{\beta}$ minimizing $\tilde{Y} - \tilde{X}'\hat{\beta}$ in the least squares sense.

This part will end with Chapter 6 which aims to apply the results to pairwise and triple preference observations using the BTL model and its extensions. Indeed, all the chosen equations presented for those models in the Section 3.4 and Chapter 4 from the previous Part, are linear in $\log w_i$ (and $\log \theta$) which allows us to write the problem as the parameters approximations of a system of linear equations. Assuming that the noise from the observations has a well-defined variance, we can thus use the results of Chapter 5 to find the BLUE of $\log \boldsymbol{w}$ in a constrained subspace. Besides, we have shown that the latter estimator minimizes the associated weighted $2-$norm over the whole space (not only the vectors minimizing the residuals as mentioned above). We eventually provide a theoretical approximation of the minimum value of the relative error on the item weights $\boldsymbol{w}$, namely $\|\boldsymbol{w} - \hat{\boldsymbol{w}}\| / \|\boldsymbol{w}\|$.

Chapter 5
General results

Consider the linear system:

$$Y = X'\beta + \epsilon \tag{5.1}$$

where:

- Y is an observation vector of dimension $m \times 1$
- X' is a known parameter matrix of dimension $m \times p$ of rank $p' \leq p$
- β is an unknown parameter vector of dimension $p \times 1$
- ϵ is an error vector of dimension $m \times 1$ with a zero mean and a variance $\mathbb{V}(\epsilon) = V$ of rank $m' \leq m$

The problem is to find the BLUE, Best Linear Unbiased Estimator $\hat{\beta}$ of β.

Definition 5.1. The Best Linear Unbiased Estimator $\hat{\beta}$ of a parameter β from a linear system $Y = X'\beta + \epsilon$ is

- a linear estimator: $\hat{\beta} = B'Y + b'_0$ with B a matrix coefficient of dimension $m \times p$ and b_0 a vector of dimension $p \times 1$
- an unbiased estimator: $\mathbb{E}[\hat{\beta}] = \beta$
- the best estimator in the sense that the variance is minimum, i.e., for any unbiased linear estimator $\tilde{\beta}$, the matrix $\mathbb{V}[\tilde{\beta}] - \mathbb{V}[\hat{\beta}]$ is positive semi-definite.

Note that linear estimator are often of the form $\hat{\beta} = B'Y$. If it is not possible to find a linear unbiased estimator of β, as it is sometimes the case, one can redefine the problem.

5.1 Gauss Theorem

Theorem 5.1.1. Consider the system (5.1) with X' of full rank i.e., $p' = p$ and the errors homoscedastic and uncorrelated, meaning that the model uncertainty is identical across all observations i.e. $\mathbb{V}[\epsilon] = \sigma^2 I$. There exists a best linear unbiased estimator of β given by

$$\hat{\beta} = \left(XX'\right)^{-1} XY \tag{5.2}$$

This Theorem has first been proven by Carl Friedrich Gauss in 1821 and then written in this modern notation by Graybill in 1961 [Gra61].

5.2 Possibly unobservable parameter β

The previous problem can also be considered for a non-full rank matrix X', i.e. $p' \leq p$. It means that there is a class of equivalence of β satisfying $Y = X'\beta + \epsilon$ and the class is defined by the set $\beta_0 + v_{X'}$ for a β_0 satisfying $Y = X'\beta_0 + \epsilon$ and $\forall v_{X'} \in \text{Ker}(X')$. Therefore, β is not observable, i.e., there is no matrix B' such that

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[B'Y] = \mathbb{E}[B'(X'\beta + \epsilon)] = B'X'\mathbb{E}[\beta] = \beta$$

since $B'X'$ can never be of full rank (and thus possibly equal to the identity matrix). The problem to find the BLUE of β is thus undefined, since the concept of bias for a class of parameters β is not defined. Two alternatives developed in the next Sections can be employed to have a consistent problem: the first one has been presented by Goldman and Zelen [GZ64] and we propose the second one as part of our contribution.

5.2.1 Sets of β's and $\hat{\beta}$'s

A first alternative, proposed in [GZ64], to this problem is to find an estimator of a function γ of β, i.e. $\gamma = L'\beta$, which is observable. In other words, there is only one vector γ such that $\gamma = L'\beta$ and $Y = X'\beta + \epsilon$. To have γ observable, each column vector of L must belong to the columnspace of X. Indeed, it comes to find $\hat{\gamma} = D'Y$ such that

$$\mathbb{E}[\hat{\gamma}] = \mathbb{E}[D'Y] = \mathbb{E}[D'(X'\beta + \epsilon)] = D'X'\mathbb{E}[\beta] = L'\beta = \gamma \iff XD = L$$

The new problem becomes to find the best linear unbiased estimator $\hat{\gamma}$ of γ for which the solution is given by the next Theorem found by Goldman and Zelen in 1964 [GZ64]:

Theorem 5.2.1. Consider the system (5.1) with X' of non-full rank, i.e. $p' \leq p$, and $\mathbb{V}[Y] = \mathbb{V}[\epsilon] = \sigma^2 I$. For any estimable function $\gamma = L'\beta$, there exists a best linear unbiased estimator of γ given by $\hat{\gamma} = L'\hat{\beta}$ where $\hat{\beta}$ is any solution to the normal equations:

$$XX'\beta = XY$$

where the solution is given by

$$\hat{\beta} = \left(XX'\right)^{-} XY + \left(I - \left(XX'\right)^{-} XX'\right) z \quad \forall z \tag{5.3}$$

where A^{-} is a generalized inverse of A [a] satisfying (a) $AA^{-}A = A$, (b) $A^{-}AA^{-} = A^{-}$ and (c) $(AA^{-})' = AA^{-}$

[a] Note that the generalized inverse differs from the Moore-Penrose pseudo-inverse since the last condition on the latter one, namely $(A^{\dagger}A)^{-1} = A^{\dagger}A$, must not necessarily be satisfied.

Since β is unobservable, it means there is a set of β's such that $Y = X'\beta + \epsilon$ defining an equivalence class. By Theorem 5.2.1, we also know there is a set of $\hat{\beta}$'s solutions of the normal equations, $XX'\hat{\beta} = XY$, i.e., the equations such that $Y - X'\hat{\beta}$ is minimum. Given any β from the set of β's and any $\hat{\beta}$ from the set of $\hat{\beta}$'s, we will generally not have $\mathbb{E}[\hat{\beta}] = \beta$ since the value of the observations $Y = X'\beta + \epsilon$ is the same for any β of the set. It holds however that

$$\forall \hat{\beta} \text{ such that } XX'\hat{\beta} = XY,\ \exists \beta \text{ such that } Y = X'\beta + \epsilon : \mathbb{E}[\hat{\beta}] = \beta$$

One may observe that for a matrix X' of full rank, in other words when there is a unique β such that $Y = X'\beta + \epsilon$, the set of $\hat{\beta}$'s reduces to a unique $\hat{\beta}$ corresponding to the solution of Theorem 5.1.1:

$$\begin{aligned}\hat{\beta} &= \left(XX'\right)^{-1} XY + \left(I - \left(XX'\right)^{-1} XX'\right) z \quad \forall z \\ &= \left(XX'\right)^{-1} XY + (I - I)z \quad \forall z \\ &= \left(XX'\right)^{-1} XY\end{aligned}$$

5.2.2 Contribution: unique β and $\hat{\beta}$ with additional constraints

A second alternative, that we propose, is to consider a new system defined by the previous one with additional conditions that β and $\hat{\beta}$ must satisfy such that both of them are unique. In that way, we restrict ourselves to a unique element of the equivalence class of β. The additional conditions on β, $S'\beta = m$ ($\beta \in C$), are defined such that β is observable, i.e. $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. Because $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$, the set of β's defining the equivalence class ($Y = X'\beta + \epsilon$) and the set of vectors $x \in C$ ($S'x = m$) have a unique intersection vector β_c satisfying $Y = X'\beta_c + \epsilon$ and $S'\beta_c = m$.

If we are interested in the estimator satisfying the same conditions, $S'\hat{\beta} = m$, the new problem is to determine, among all the vectors belonging to C, the BLUE estimator of β_c satisfying:

$$Y = X'\beta_c + \epsilon \quad m = S'\beta_c \quad \text{with } \text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p,\ \mathbb{E}(Y) = X'\beta_c,\ \mathbb{V}(Y) = \sigma^2 I$$

Let Q' be an orthonormal basis of $\text{Ker}(S')$ satisfying $QQ' = I$. The following Lemma presents a useful property of the matrix Q for future results:

Lemma 5.2.1. *Let X' be a matrix of dimension $m \times p$ of rank $p' \leq p$. Let S be a matrix of dimension $p \times (p - p')$ such that $\begin{bmatrix} X & S \end{bmatrix}$ is full rank and let Q' be an orthonormal basis* $\text{Ker}(S')$. *The matrix $X'_p = X'Q'$ is full rank.*

Proof. As the matrix Q' is orthonormal, it is full rank so $\text{rank}(Q') = \text{rank}(Q) = p'$ and $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. By the definition of the rank,

$$\text{rank}\left(Q\begin{bmatrix} X & S \end{bmatrix}\right) \leq \min\left\{\text{rank}(Q), \text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right)\right\} = \min\left\{p', p\right\} = p'$$

The inequality of Sylvester provides a lower bound on the rank of the matrix $Q\begin{bmatrix} X & S \end{bmatrix}$:

$$\text{rank}\left(Q\begin{bmatrix} X & S \end{bmatrix}\right) \geq \text{rank}(Q) + \text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) - p = p'$$

The rank of the matrix $Q\begin{bmatrix} X & S \end{bmatrix}$ is thus equal to p'.

Because $Q' = \begin{bmatrix} q_1 & \cdots & q_{p'} \end{bmatrix}$ is a basis of $\mathrm{Ker}(S')$, the columns of the matrix belong to the space implying that $S'q_i = 0$. Therefore $S'Q' = 0$ and

$$Q \begin{bmatrix} X & S \end{bmatrix} = \begin{bmatrix} QX & QS \end{bmatrix} = \begin{bmatrix} QX & 0 \end{bmatrix}$$

The rank of the matrix $Q \begin{bmatrix} X & S \end{bmatrix}$ is thus equal to the rank of the matrix QX and finally as $\mathrm{rank}(QX) = p'$, the matrix is full rank. □

Thanks to this Lemma, it is now possible to find the BLUE estimator of our new problem with homogeneous additional constraints of the type $S'\beta = 0$. This is explained in the following Lemma:

Lemma 5.2.2. *Consider the system* (5.1) *with* $\mathbb{V}[Y] = \mathbb{V}[\epsilon] = \sigma^2 I$. *Let* S *be a matrix of size* $p \times (p - p')$ *such that* $\mathrm{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. *Among the possible vectors* β *of the equivalence class defined by* $Y = X'\beta + \epsilon$, *the vector* β_c *of that class which respects the additional constraints* $S'\beta_c = 0$ *is unique. Let* Q' *be an orthonormal basis of* $\mathrm{Ker}(S')$. *Then there exists a best linear unbiased estimator of* β_c *in the subspace* $C = \{x \in \mathbb{R}^p | S'x = 0\}$ *which is given by*

$$\hat{\beta}_c = \left(Q'QXX'Q'Q\right)^\dagger XY \tag{5.4}$$

Proof. The initial system $Y = X'\beta + \epsilon$ is satisfied by a set of different β defining an equivalence class. This set is defined by p' independent equations, i.e. $\mathrm{rank}(X') = p'$. The conditions $S'x = 0$ found by adding $p - p'$ independent equations, i.e. $\mathrm{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$, to the previous system, allows us to define a unique vector β_c satisfying $Y = X'\beta_c + \epsilon$ and $S'\beta_c = 0$. Using $X'_p = X'Q'$ and b_p, the coordinates of β_c in the basis Q', the problem can thus be reformulated as:

$$Y = X'\beta_c + \epsilon = X'(Q'b_p) + \epsilon = X'_p b_p + \epsilon$$

Since X'_p is full rank by Lemma 5.2.1, Theorem 5.1.1 can be used, and the best linear unbiased estimator of b_p is given by

$$\hat{b}_p = \left(X_p X'_p\right)^{-1} X_p Y = \left(QXX'Q'\right)^{-1} QXY$$

This value can be replaced in the definition of β_c. Using the properties of the Moore-Penrose pseudo-inverse and knowing that $Q' = Q^\dagger$ the solution becomes:

$$\begin{aligned} \hat{\beta}_c &= Q'\hat{b}_p = Q'\left(QXX'Q'\right)^{-1} QXY \\ &= Q^\dagger \left(QXX'Q'\right)^\dagger (Q')^\dagger XY = \left(Q'QXX'Q'Q\right)^\dagger XY \end{aligned}$$[1]

It remains to show that $\hat{\beta}_c$ is the BLUE estimator of β_c in the subspace C. The vector is linear since it can be written in the form $\hat{\beta}_c = D'Y$ with

[1] This last expression is obtained using a property of the Moore-Penrose pseudo-inverse, namely $(AB)^\dagger = B^\dagger A^\dagger$ when A has orthonormal columns ($A'A = A^\dagger A = I$) or when B has orthonormal rows ($BB' = BB^\dagger = I$) and since Q has orthonormal rows, we obtain the above result.

$D' = (Q'QXX'Q'Q)^{\dagger} X$. It can also be proven that it is unbiased. As $\hat{b}_p$ is the best linear unbiased estimator of b_p,

$$\mathbb{E}[\hat{\beta}_c] = Q'\mathbb{E}[\hat{b}_p] = Q'b_p = \beta_c$$

To be the BLUE of β_c, $\hat{\beta}_c$ must also minimize its variance namely for any unbiased linear estimator $\tilde{\beta}_c$, the matrix $\mathbb{V}(\tilde{\beta}_c) - \mathbb{V}(\hat{\beta}_c)$ must be positive semi-definite. For all vectors $\tilde{\beta}_c$ belonging to C, there is a vector

$$\begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix} = \begin{bmatrix} Q \\ 0_{(p-p')\times p} \end{bmatrix} \tilde{\beta}_c \quad \text{such that} \quad \tilde{\beta}_c = \begin{bmatrix} Q' & 0_{p\times(p-p')} \end{bmatrix} \begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix}$$

Because $\hat{b}_p$ is the BLUE of b_p, this vector satisfies

$$\mathbb{V}\left(\begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) - \mathbb{V}\left(\begin{bmatrix} \hat{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) = \begin{bmatrix} \mathbb{V}(\tilde{b}_p) - \mathbb{V}(\hat{b}_p) & 0_{p'\times(p-p')} \\ 0_{(p-p')\times p'} & 0_{(p-p')\times(p-p')} \end{bmatrix} \succeq 0 \tag{5.5}$$

Using this vector, it is also true that

$$\begin{aligned} \mathbb{V}(\tilde{\beta}_c) - \mathbb{V}(\hat{\beta}_c) &= \mathbb{V}\left(\begin{bmatrix} Q' & 0_{p\times(p-p')} \end{bmatrix} \begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) - \mathbb{V}\left(\begin{bmatrix} Q' & 0_{p\times(p-p')} \end{bmatrix} \begin{bmatrix} \hat{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) \\ &= \begin{bmatrix} Q' & 0_{p\times(p-p')} \end{bmatrix} \left(\mathbb{V}\begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) \begin{bmatrix} Q \\ 0_{(p-p')\times p} \end{bmatrix} \end{aligned}$$

To be positive semi-definite, the following relation must be satisfied:

$$\forall x \in \mathbb{R}^p,\ x' \begin{bmatrix} Q' & 0_{p\times(p-p')} \end{bmatrix} \left(\mathbb{V}\begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) \begin{bmatrix} Q \\ 0_{(p-p')\times p} \end{bmatrix} x \geq 0$$

which can be reformulated as

$$\forall \tilde{y} \in \mathbb{R}^p \text{ such that } \tilde{y} = \begin{bmatrix} Q \\ 0_{(p-p')\times p} \end{bmatrix} x,\ \tilde{y}' \left(\mathbb{V}\begin{bmatrix} \tilde{b}_p \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{b}_p \\ 0_{(p-p')} \end{bmatrix}\right) \tilde{y} \geq 0$$

Using the relation (5.5), this relation is always satisfied. □

In the above Lemma, we restricted ourselves to the case of additional constraints of the form $S'\beta = 0$. This result can be generalized to problems with any type of constraints $S'\beta = m$ which is explained in the following Theorem:

Theorem 5.2.2. Consider the system (5.1) with $\mathbb{V}[Y] = \mathbb{V}[\epsilon] = \sigma^2 I$. Let S be a matrix of size $p \times (p - p')$ such that $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. Among the possible vectors β of the equivalence class defined by $Y = X'\beta + \epsilon$, the vector β_c of that class which respects the additional constraints $S'\beta_c = m$ is unique. Let Q' be an orthonormal basis of $\text{Ker}(S')$. Then there exits a best linear unbiased estimator of β_c in the subspace $C = \{x \in \mathbb{R}^p | S'x = m\}$ which is given by

$$\hat{\beta}_c = \left(Q'QXX'Q'Q\right)^{\dagger} XY + B(S'B)^{-1}m \tag{5.6}$$

where B is a basis of $\text{Ker}(X')$. This vector also satisfies

$$XY = XX'\hat{\beta}_c$$

Proof. Because β_c belongs to a class of equivalence satisfying $Y = X'\beta + \epsilon$, we know that the vector is defined by $\beta_c = \beta_0 + v_{X'}$ where β_0 is an element of that class and $v_{X'} \in \text{Ker}(X')$. Define β_0 as the vector satisfying $Y = X'\beta_0 + \epsilon$ and $S'\beta_0 = 0$. Then the constraints are such that $S'\beta_c = S'(\beta_0 + v_{X'}) = S'v_{X'} = m$. Denote by B a basis of $\text{Ker}(X')$, then $v_{X'} = \sum_{i=1}^{p-p'} B_i\alpha_i = B\alpha$. The constraints become $S'B\alpha = m$ and because S' and B are full rank matrices, we obtain $\alpha = (S'B)^{-1}m$. In other words, the vector one wants to approximate can be seen as $\beta_c = \beta_0 + B(S'B)^{-1}m$.

From Lemma 5.2.2, we know that the BLUE of β_0 is given by

$$\hat{\beta}_0 = \left(Q'QXX'Q'Q\right)^{\dagger} XY$$

This vector satisfies

$$\begin{aligned} QXX'\hat{\beta}_0 &= QXX'\left(Q'QXX'Q'Q\right)^{\dagger} XY \\ &= QXX'Q'\left(QXX'Q'\right)^{-1} QXY \ ^{2} = QXY \end{aligned}$$

The above expression can be written as

$$XX'\hat{\beta}_0 = XY + q$$

where $q \in \text{Ker}(Q)$ since $Q(XX'\hat{\beta}_0 - XY) = 0$. We know furthermore S forms a basis of $\text{Ker}(Q)$ since Q' is an orthonormal basis of $\text{Ker}(S')$. This means that $q \in \text{Ker}(Q)$ can be written as $S\alpha$ for all vector α of dimension $p - p'$ and the above expression becomes

$$XX'\hat{\beta}_0 = XY + S\alpha \qquad \forall\alpha \in \mathbb{R}^{p-p'}$$

We also know that the matrix S must be chosen such that $\begin{bmatrix} X & S \end{bmatrix}$ is full rank. In other words, for all vector v belonging to $\text{Ker}(X')$, it must hold that $S'v = 0$. Pre-multiplying the above expression by $v \in \text{Ker}(X')$ gives

$$v'XX'\hat{\beta}_0 = v'XY + v'S\alpha \qquad \forall\alpha \in \mathbb{R}^{p-p'}$$

for which α must be equal to zero since $v'X = 0$. We have thus shown that the BLUE of β_0 satisfies the following expression:

$$XX'\hat{\beta}_0 = XY$$

In other words, it means that for all the particular vector β_0^p belonging to a constrained subspace $S'\beta_0 = 0$ (thus considering all the possible constraints of

[2]Using the property of the Moore-Penrose pseudo-inverse stating that $(AB)^{\dagger} = B^{\dagger}A^{\dagger}$ if A has orthonormal columns ($A'A = A^{\dagger}A = I$) or if B has orthonormal rows ($BB' = BB^{\dagger} = I$) and because Q has orthonormal rows.

the form $S'\beta_0 = 0$), the BLUE of β_0^p is such that $\hat{\beta}_0^p = \hat{\beta}_0 + v_{XX'}$ where $\hat{\beta}_0$ is a vector satisfying $XX'\hat{\beta}_0 = XY$ and $v_{XX'} \in \text{Ker}(XX')$. Actually, we have that $v_{XX'} \in \text{Ker}(X')$ since $\text{Ker}(XX') = \text{Ker}(X')$:

$$\begin{aligned} \forall v \in \text{Ker}(X') : X'v = 0 &\implies XX'v = 0 \implies v \in \text{Ker}(XX') \\ \forall v \in \text{Ker}(XX') : XX'v = 0 &\implies v'XX'v = (X'v)'X'v = 0 \\ &\implies X'v = 0 \implies v \in \text{Ker}(X') \end{aligned}$$

We are looking for an unbiased estimator of $\beta_c = \beta_0 + B(S'B)^{-1}$. To do so, let us apply the same translation $B(S'B)^{-1}$ to the BLUE of β_0:

$$\hat{\beta}_c = \left(Q'QXX'Q'Q\right)^{\dagger} XY + B(S'B)^{-1}m$$

This estimator is the BLUE of β_c. We indeed have that $\hat{\beta}_c$ is a linear and unbiased estimator of β_c since it is a linear combination of the observations to which a constant vector $v_{X'}$ is added, and it holds that

$$\mathbb{E}[\hat{\beta}_c] = \mathbb{E}[\hat{\beta}_0 + B(S'B)^{-1}m] = \beta_0 + B(S'B)^{-1}m = \beta_c$$

since $\mathbb{E}[\hat{\beta}_0] = \beta_0$ as $\hat{\beta}_0$ is an unbiased estimator of β_0.

We still have to prove that the found estimator is the best one in the sense that it minimizes the variance. Take $\tilde{\beta}_c$, another unbiased linear estimator of β_c. Then $\tilde{\beta}_c$ can be written as $\tilde{\beta}_c = \tilde{\beta}_0 + \tilde{v}_{X'}$ where $\tilde{\beta}_0$ is a linear unbiased estimator of β_0 and $\tilde{v}_{X'} \in \text{Ker}(X')$. Because $\hat{\beta}_0$ is the BLUE of β_0 it holds that

$$\mathbb{V}(\hat{\beta}_c) = \mathbb{V}(\hat{\beta}_0 + v_{X'}) = \mathbb{V}(\hat{\beta}_0) \preceq \mathbb{V}(\tilde{\beta}_0) = \mathbb{V}(\tilde{\beta}_c)$$

It holds furthermore that

$$\begin{aligned} XX'\hat{\beta}_c &= XX'\left(\hat{\beta}_0 + B(S'B)^{-1}m\right) \\ &= XX'\hat{\beta}_0 + XX'B(S'B)^{-1}m \\ &= XY \end{aligned}$$

since $XX'\hat{\beta}_0 = XY$ and $B(S'B)^{-1}m \in \text{Ker}(X')$. □

One may again observe that for a matrix X' of full rank, i.e., a unique β such that $Y = X'\beta + \epsilon$, the subspace $C = \{x \in \mathbb{R}^p | S'x = m\}$ such that $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$ is $\mathbb{R}^p$ since S must be the null matrix. Therefore, in that case, the estimator corresponds to the solution of Theorem 5.1.1:

$$\hat{\beta} = \left(XX'\right)^{-1} XY \quad (Q'QX = X)$$

Note that it is also possible to find an expression of the BLUE of β_c by using Theorem 5.2.1, and it is detailed in Appendix C. This expression is actually difficult to use in practice as to find the unique solution to the problem, one must find the corresponding matrix among the generalized inverse of XX' and the unknown vector z and the method is not straightforward since it requires computing $\hat{\gamma}$ to find $\hat{\beta}_c$.

5.3 Possibly non-independent and non-homoscedastic observations

In the previous Section 5.2, we have shown how to redefine the problem of finding the Best Linear Unbiased Estimator of parameters β satisfying $Y = X'\beta + \epsilon$ with a non-full rank matrix X'. This problem still consider an error vector with $\mathbb{V}[\epsilon] = \sigma I$ as in Theorem 5.1.1 proven by Gauss and presented in Section 5.1. However, in some applications, errors might be such that $\mathbb{V}[\epsilon] = V$ with $\text{rank}(V) = m' \leq m$, i.e., observations might be non-independent, correlated or with different variances.

The number of observations and equations can be reduced without losing any information and the variance of the observations can be normalized to have again the error variance equal to σI. This gives rise to the following Theorem whose general idea was first introduced by Goldman and Zelen [GZ64] but for which explanation is reformulated in a more straightforward way.

Theorem 5.3.1. Consider the system (5.1) with X' possibly of non-full rank and $\mathbb{V}[\epsilon] = V$ where V can be singular, i.e. $\text{rank}(V) = m' \leq m$. This system can be reformulated as

$$\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon} \quad \text{with} \quad K'\beta = l$$

where:

- $\tilde{Y} = T'Y$ is a redefined observation vector of dimension $m' \times 1$
- $\tilde{X} = XT$ is a redefined known parameter matrix of dimension $m' \times p$ of rank $p' \leq p$
- β is the unknown parameter vector of dimension $p \times 1$
- $\tilde{\epsilon} = T'\epsilon$ is an error vector of dimension $m' \times 1$ with a zero mean and a variance $\mathbb{V}(\tilde{\epsilon}) = I$
- $K = XF_2$ and $l = F_2'Y$

with T such that $TT' = V^\dagger$ and $F = \begin{bmatrix} F_1 & F_2 \end{bmatrix}$, any orthonormal basis of $\text{Ker}(V)$ such that $\text{rank}(F_1'X') = 0$ and $F_2'X'$ of full rank.

Proof. Due to the fact that V is non-full rank, there exists a matrix F of dimension $m \times (m - m')$ satisfying $F'V = 0$ (and thus $F'Y$ being constant as $\mathbb{V}[F'Y] = F'VF = 0$). The matrix F is an orthogonal basis of $\text{Ker}(V)$ and can be chosen orthonormal, i.e. $F'F = I$.

As it is explained in [GZ64], $F'X'$ will generally be equal to 0 but in some cases it won't be. They define thus $F = \begin{bmatrix} F_1 & F_2 \end{bmatrix}$ with F_i of size $m \times s_i$ such that

$$\text{rank}(F_1'X') = 0 \quad \text{and} \quad \text{rank}(F_2'X') = s_2 \text{ for } F'V = 0$$

Because any covariance matrix is symmetric positive semi-definite, its singular value decomposition (SVD) has the form

$$P'VP = \begin{bmatrix} 0_{(m-m')\times(m-m')} & 0_{(m-m')\times m} \\ 0_{m\times(m-m')} & \Lambda \end{bmatrix}$$

The matrix P can thus be decomposed as $P = \begin{bmatrix} F & G \end{bmatrix}$ since $F'V = VF = 0$. This gives

$$V = P \begin{bmatrix} 0 & 0 \\ 0 & \Lambda \end{bmatrix} P' = G\Lambda G' \tag{5.7}$$

The equation of the problem can be written as

$$\begin{aligned} P'Y &= P'X'\beta + P'\epsilon \\ \begin{bmatrix} F'Y \\ G'Y \end{bmatrix} &= \begin{bmatrix} F'X'\beta + F'\epsilon \\ G'X'\beta + F'\epsilon \end{bmatrix} \end{aligned} \tag{5.8}$$

where P is derived from the SVD of V as before. As $F'Y'$ is constant, the first equations can be written as $F'Y = F'\mathbb{E}[Y] = F'X'\beta$. The division of F in two block matrices is thus important as it results in

$$F_1'Y' = F_1'X'\beta = 0 \quad \text{and} \quad F_2'Y' = F_2'X'\beta$$

The left-side equations do not provide any additional information. However, if $F_2 \neq 0$, the problem contains in that case additional constraints on β that must be satisfied, namely $F_2'Y = F_2'X'\beta$. The problem reduces to the second equations of the reformulation (5.8) under those conditions :

$$G'Y = G'X'\beta + G'\epsilon \quad \text{such that} \quad F_2'Y = F_2'X'\beta \tag{5.9}$$

It remains to have a noise variance equal to the identity matrix. By the definition of the pseudo-inverse and the decomposition of P:

$$V^\dagger = P \begin{bmatrix} 0_{(m-m')\times(m-m')} & 0_{(m-m')\times m} \\ 0_{m\times(m-m')} & \Lambda^{-1} \end{bmatrix} P' = G\Lambda^{-1}G'$$

Let us define $T = G\Lambda^{-1/2}$. To end with a problem having a noise variance equal to the identity matrix, the equation (5.9) can be pre-multiplied by $\Lambda^{-1/2}$ leading to

$$\tilde{Y} = \tilde{X}\beta + \tilde{\epsilon} \quad \text{such that} \quad K'\beta = l$$

where $\tilde{Y} = T'Y$, $\tilde{X} = XT$ and $\tilde{\epsilon} = T'\epsilon$ and $T = G\Lambda^{-1/2}$. It can be proven that the noise of this new problem has a zero mean and an identity variance. It indeed holds that $\mathbb{E}[\tilde{\epsilon}] = T'\mathbb{E}[\epsilon] = 0$.

The variance immediately stems from the definition of T:

$$\mathbb{V}[\tilde{\epsilon}] = T'\mathbb{V}[\epsilon]T = T'VT = \Lambda^{-1/2}G'G\Lambda G'G\Lambda^{-1/2}I$$

□

When $\text{Ker}(V) \subseteq \text{Ker}(X)$, Theorem 5.3.1 permits us to reduce our problem to the case of Section 5.2, i.e., $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ with $\mathbb{V}[\tilde{\epsilon}] = I$ without any other condition.

It allows us to extend Theorems 5.2.1 and 5.2.2 to the general case for any variance $\mathbb{V}[\epsilon] = V$ possibly singular with Theorem 5.3.2 and 5.3.3 respectively. For the case $\text{Ker}(V) \not\subseteq \text{Ker}(X)$, i.e., with conditions $K'\beta = m$, refer to Gauss-Markov Theorem with given restraints in [GZ64].

Theorem 5.3.2. General case of Theorem 5.2.1: Consider the system (5.1) with X' possibly of non-full rank, i.e. $\text{rank}(X') = p' \leq p$, and $\mathbb{V}[\epsilon] = V$ possibly singular, i.e. $\text{rank}(V) = m' \leq m$. If $\text{Ker}(V) \subseteq \text{Ker}(X)$, the best linear unbiased estimator of any estimable function $\gamma = L'\beta$ exists and is given by $\hat{\gamma} = L'\hat{\beta}$ where $\hat{\beta}$ is any solution to the normal equations:

$$XV^{\dagger}X'\beta = XV^{\dagger}Y$$

where the solution is given by

$$\hat{\beta} = \left(XV^{\dagger}X'\right)^{-} XV^{\dagger}Y + \left(I - \left(XV^{\dagger}X'\right)^{-} XV^{\dagger}X'\right) z \quad \forall z \tag{5.10}$$

with A^{-}, a generalized inverse of A satisfying (a) $AA^{-}A = A$, (b) $A^{-}AA^{-} = A^{-}$ and (c) $(AA^{-})' = AA^{-}$ and $A^{\dagger}$, the unique pseudoinverse of A.

Proof. By Theorem 5.3.1, the problem can be reformulated as

$$\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$$

where $\tilde{Y} = T'Y$, $\tilde{X} = XT$, $\tilde{\epsilon} = T'\epsilon$ and T defined as in the Theorem.

Since $\mathbb{V}[\tilde{Y}] = \mathbb{V}[\tilde{\epsilon}] = I$, Theorem 5.2.1 can be used with $\tilde{Y}, \tilde{X}, \tilde{\epsilon}$ instead of Y, X, ϵ and knowing that $TT' = G\Lambda^{-1/2}\Lambda^{-1/2}G' = V^{\dagger}$. □

Theorem 5.3.3. General case of Theorem 5.2.2: Consider the system (5.1) with X' possibly of non-full rank, i.e. $\text{rank}(X') = p' \leq p$, and $\mathbb{V}[\epsilon] = V$ possibly singular, i.e. $\text{rank}(V) = m' \leq m$. Among the possible vectors β, denote β_c as the unique one satisfying the additional conditions $S'\beta = m$ for any matrix S of size $p \times (p - p')$ such that $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. Let Q' be an orthonormal basis of $\text{Ker}(S')$. If $\text{Ker}(V) \subseteq \text{Ker}(X)$, the best linear unbiased estimator of β_c from this redefined system exists and is given by:

$$\hat{\beta}_c = \left(Q'QXV^{\dagger}X'Q'Q\right)^{\dagger} XV^{\dagger}Y + B(S'B)^{-1}m$$

where B is a basis of $\text{Ker}(X')$. This vector also satisfies

$$XV^{\dagger}Y = XV^{\dagger}X'\hat{\beta}_c$$

Proof. By Theorem 5.3.1, for $\text{Ker}(X) = \text{Ker}(V)$, the problem can be reformulated as

$$\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$$

where $\tilde{Y} = T'Y$, $\tilde{X} = XT$, $\tilde{\epsilon} = T'\epsilon$ and T defined as in the Theorem.

Since $\mathbb{V}[\tilde{Y}] = \mathbb{V}[\tilde{\epsilon}] = I$, Theorem 5.2.2 can be used if $\text{rank}\left(\begin{bmatrix}\tilde{X} & S\end{bmatrix}\right) = p$. Note that:

$$rank(XT) = rank(XG) = rank(XP) = rank(X) = p'$$

since $XF = 0$ and P is an orthonormal matrix $\begin{bmatrix}F & G\end{bmatrix}$ defined by (5.7).

If the columns of $\tilde{X}$ are linearly independent of those of S, $\text{rank}\left(\begin{bmatrix}\tilde{X} & S\end{bmatrix}\right) = rank(\tilde{X}) + rank(S) = p' + (p - p') = p$. Linear independence between columns is equivalent to:

$$\nexists\ \alpha_i : s_i = \tilde{X}\alpha_i = XT\alpha_i \quad \text{for } S = \begin{bmatrix}s_1 & \cdots & s_{(p-p')}\end{bmatrix}$$

which can be reformulated with $\gamma_i = T\alpha_i$, i.e., γ_i belong to the columnspace of T:

$$\nexists\ \gamma_i \in \mathscr{C}(T) : s_i = X\gamma_i \quad \text{for } S = \begin{bmatrix}s_1 & \cdots & s_{(p-p')}\end{bmatrix}$$

which holds since the columns of X are linearly independent of those of S by assumption. Thus, Theorem 5.2.2 can be used with $\tilde{Y}$, $\tilde{X}$, $\tilde{\epsilon}$ instead of Y, X, ϵ and knowing that $TT' = V^\dagger$:

$$\begin{aligned}\hat{\beta}_c &= \left(Q'Q\tilde{X}\tilde{X}'Q'Q\right)^\dagger \tilde{X}\tilde{Y} + B(S'B)^{-1}m \\ &= \left(Q'QXV^\dagger X'Q'Q\right)^\dagger XV^\dagger Y + B(S'B)^{-1}m\end{aligned}$$

From Theorem 5.2.2, we know that this estimator satisfies

$$\tilde{X}\tilde{X}'\hat{\beta}_c = \tilde{X}\tilde{Y}$$

and replacing $\tilde{X}$ and $\tilde{Y}$ by their definition we obtain

$$XV^\dagger X'\hat{\beta}_c = XV^\dagger Y$$

□

The specific result from Theorem 5.2.2 can be obtained for $V = \sigma^2 I$, since in that case, by the definition of the Moore-Penrose pseudo-inverse, $V^\dagger = V^{-1} = \dfrac{1}{\sigma^2}I$:

$$\hat{\beta}_c = \left(Q'QX\frac{1}{\sigma^2}X'Q'Q\right)^\dagger X\frac{1}{\sigma^2}Y + B(S'B)^{-1}m = \left(Q'QXX'Q'Q\right)^\dagger XY + B(S'B)^{-1}m$$

Note that the above Theorem extend Theorem 5.2.2 for which β becomes observable by imposing mandatory condition $S'\beta = m$ and using the knowledge of those conditions before solving the least squares problem (alternative 2). One could also be interested in the problem of finding the unique estimator respecting those conditions by considering the solution of Theorem 5.2.1 (alternative 1). The extension of Theorem C.1.1 is given in Appendix C.

5.3.1 Discussion of the additional conditions on β and $\hat{\beta}$

In the initial problem, i.e., without adding supplementary conditions on β, we know that we have a set of β's, such that $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ were $\tilde{Y} = T'Y$, $\tilde{X} = XT$, $\tilde{\epsilon} = T'\epsilon$ and T defined as in the Theorem 5.3.1 with $TT' = V^\dagger$. From that latter Theorem, we also know that for each particular β^p of that set, its BLUE $\hat{\beta}_c^p$ belongs to a set defined by $XV^\dagger X'\hat{\beta} = XV^\dagger Y$. In other words, any $\hat{\beta}$ minimizes $\tilde{Y} - \tilde{X}'\hat{\beta}$ in the least squares sense. Those two sets ($\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ and $XV^\dagger X'\hat{\beta} = XV^\dagger Y$) are in fact parallel since $\tilde{X}\tilde{X}' = XV^\dagger X'$ and $\text{Ker}(\tilde{X}') = \text{Ker}(\tilde{X}\tilde{X}')$:

$$\begin{aligned} \forall v \in \text{Ker}(\tilde{X}') : \tilde{X}'v = 0 &\implies \tilde{X}\tilde{X}'v = 0 \implies v \in \text{Ker}(\tilde{X}\tilde{X}') \\ \forall v \in \text{Ker}(\tilde{X}\tilde{X}') : \tilde{X}\tilde{X}'v = 0 &\implies v'\tilde{X}\tilde{X}'v = (\tilde{X}'v)'\tilde{X}'v = 0 \\ &\implies \tilde{X}'v = 0 \implies v \in \text{Ker}(\tilde{X}') \end{aligned} \tag{5.11}$$

In order to have a unique vector β^p from the set $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$, we defined a new system with additional conditions on β^p:

$$Y = X'\beta^p + \epsilon \quad m = S'\beta^p \quad \textit{with rank} \left(\begin{bmatrix} X & S \end{bmatrix} \right) = p,\ \mathbb{E}(Y) = X'\beta^p,\ \mathbb{V}(Y) = \sigma^2 I$$

However, this does not make the estimator of β^p unique since there exists an infinite number of constraints satisfied by β^p as shown on Figure 5.1 for two distinct constraints ($\beta_{c1}^p = \beta_{c2}^p = \beta^p$). Nevertheless, it holds that in a constrained subspace $C = \{x \in \mathbb{R}^p | S'x = m\}$ such that $\beta^p \in C$, the vector has a BLUE $\hat{\beta}_c^p$ given by Theorem 5.3.3. Actually, by choosing the additional constraints the estimator of β^p must satisfy, a particular weighted 2−norm defined by $\left\|\beta^p - \hat{\beta}\right\|_{2,A} = \sqrt{(\beta^p - \hat{\beta})'A(\beta^p - \hat{\beta})}$ with A a positive definite matrix is minimized among the estimators minimizing $\tilde{Y} - \tilde{X}'\hat{\beta}$ in the least squares sense.

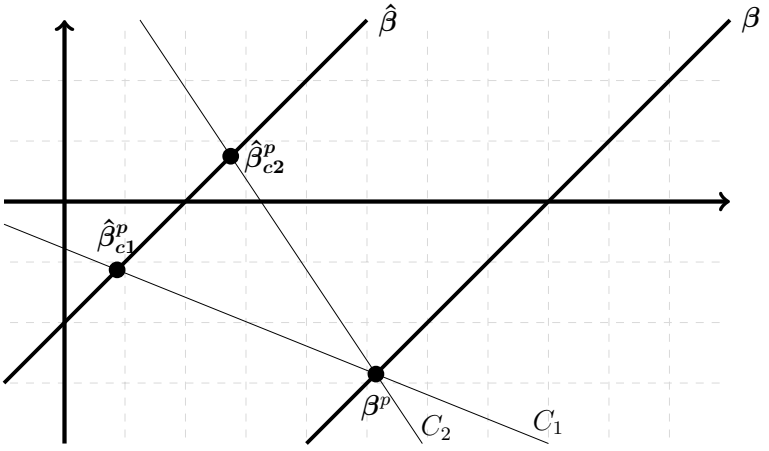


Figure 5.1: Illustration of the fact that for a particular β^p satisfying $\tilde{Y} = \tilde{X}'\beta^p + \tilde{\epsilon}$, several constraints of the form $S'\beta = m$ can be considered. Regarding those constraints, the estimator of β^p will differ. The example considers two parameters and one additional constraint needed to render the parameters observable.

The following Theorem links the constraints added to render β unique and the weighted 2−norm minimized:

Theorem 5.3.4. Consider the system (5.1) and the Theorem 5.3.3. If the additional conditions $S'\beta = m$ (for β^p and $\hat{\beta}_c^p$) are such that:

$$\forall x \in \mathrm{Ker}(X'),\ \forall s \in \mathrm{Ker}(S') : \quad \langle x, s\rangle_A = x'As = 0$$

for A symmetric positive definite or in other words, if $\mathrm{Ker}(S')$ is the orthogonal complement of $\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_A$, then:

$$\hat{\beta}_c^p = \underset{XV^\dagger X'\hat{\beta}=XV^\dagger Y}{\arg\min} \left\|\beta^p - \hat{\beta}\right\|_{2,A} = \underset{XV^\dagger X'\hat{\beta}=XV^\dagger Y}{\arg\min} \sqrt{(\beta^p - \hat{\beta})'A(\beta^p - \hat{\beta})}$$

for the induced norm $\|\cdot\|_A = \sqrt{\langle\cdot,\cdot\rangle_A}$

Proof. By Theorem 5.3.1, we know that the set of β's is the set such that $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ where $\tilde{Y} = T'Y$, $\tilde{X} = XT$, $\tilde{\epsilon} = T'\epsilon$ and T defined as in the Theorem with $TT' = V^\dagger$. It holds furthermore that $\mathrm{Ker}(\tilde{X}') = \mathrm{Ker}(X')$ since $\mathrm{Ker}(X') \in \mathrm{Ker}(\tilde{X}') = \mathrm{Ker}(T'X')$ and $\mathrm{rank}(XT) = rank(XG) = rank(XP) = rank(X) = p'$ with P and G defined by (5.7). This means that any β satisfying $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ can be represented by

$$\beta = \beta_0 + v_{\tilde{X}'} = \beta_0 + v_{X'}$$

for $\tilde{Y} = \tilde{X}'\beta_0 + \tilde{\epsilon}$ and $v_{\tilde{X}'} \in \mathrm{Ker}(\tilde{X}'), v_{X'} \in \mathrm{Ker}(X')$.

We also know from Theorem 5.3.3 that $\hat{\beta}$'s is the set such that $XV^\dagger Y = XV^\dagger X'\hat{\beta}$. It holds that $\mathrm{Ker}(X') = \mathrm{Ker}(XV^\dagger X')$ since

$$\begin{aligned}
\forall v \in \mathrm{Ker}(\tilde{X}') : \tilde{X}'v = 0 &\implies \tilde{X}\tilde{X}'v = 0 \implies v \in \mathrm{Ker}(\tilde{X}\tilde{X}') \\
\forall v \in \mathrm{Ker}(\tilde{X}\tilde{X}') : \tilde{X}\tilde{X}'v = 0 &\implies v'\tilde{X}\tilde{X}'v = (\tilde{X}'v)'\tilde{X}'v = 0 \\
&\implies \tilde{X}'v = 0 \implies v \in \mathrm{Ker}(\tilde{X}')
\end{aligned}$$

and because $\tilde{X}\tilde{X}' = XV^\dagger X'$ and $\mathrm{Ker}(X') = \mathrm{Ker}(\tilde{X}')$. This means that any $\hat{\beta}$ satisfying $XV^\dagger Y = XV^\dagger X'\hat{\beta}$ can be represented by

$$\hat{\beta} = \hat{\beta}_0 + v_{XV^\dagger X'} = \hat{\beta}_0 + v_{X'} \tag{5.12}$$

for $XV^\dagger Y = XV^\dagger X'\hat{\beta}_0$ and $v_{XV^\dagger X'} \in \mathrm{Ker}(XV^\dagger X'), v_{X'} \in \mathrm{Ker}(X')$.

Let us define $\hat{\beta}_{c2}^p \neq \hat{\beta}_c^p$, any $\hat{\beta}$ different from $\hat{\beta}_c^p$ such that $XV^\dagger Y = XV^\dagger X'\hat{\beta}$. Thanks to equation (5.12), it can be written as:

$$\begin{aligned}
\beta^p &= \beta_0 + v_{X'} \\
\hat{\beta}_c^p &= \hat{\beta}_0 + v_{X'} \\
\hat{\beta}_{c2}^p &= \hat{\beta}_c + \tilde{v}_{X'}
\end{aligned}$$

with $\hat{\beta}_0, \beta_0 \in \mathrm{Ker}(S')$ and $\tilde{v}_{X'}, v_{X'} \in \mathrm{Ker}(X')$ (where $v_{X'} = B(S'B)^{-1}m$ as in Theorem 5.3.3).

We can develop the square of the norm for $\hat{\beta}_d$ since minimizing the square of a norm is equivalent to minimize the norm:

$$\begin{aligned}
\left\|\beta^p - \hat{\beta}^p_{c2}\right\|^2_{2,A} &= \left\|\beta^p - \left(\hat{\beta}^p_c + \tilde{v}_{X'}\right)\right\|^2_{2,A} \\
&= \left\|\beta^p - \hat{\beta}^p_c\right\|^2_{2,A} + \|\tilde{v}_{X'}\|^2_{2,A} - 2\left\langle \beta^p - \hat{\beta}^p_c, \tilde{v}_{X'}\right\rangle_A \\
&= \left\|\beta^p - \hat{\beta}^p_c\right\|^2_{2,A} + \|\tilde{v}_{X'}\|^2_{2,A} - 2\left\langle \beta_0 - \hat{\beta}_0, \tilde{v}_{X'}\right\rangle_A \\
&= \left\|\beta^p - \hat{\beta}^p_c\right\|^2_{2,A} + \|\tilde{v}_{X'}\|^2_{2,A} \\
&> \left\|\beta^p - \hat{\beta}^p_c\right\|^2_{2,A} \text{ for } \tilde{v}_{X'} = \hat{\beta}^p_{c2} - \hat{\beta}^p_c \neq \mathbf{0}
\end{aligned}$$

In the above computations, we used the fact that $\left\langle \beta_0 - \hat{\beta}_0, \tilde{v}_{X'}\right\rangle_A = 0$ since $\beta_0 - \hat{\beta}_0 \in \mathrm{Ker}(S')$ and $\tilde{v}_{X'} \in \mathrm{Ker}(X')$ and $\mathrm{Ker}(S')$ is the orthogonal complement of $\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_A$. It proves that if $\mathrm{Ker}(S')$ is the orthogonal complement of $\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_A$, $\hat{\beta}^p_c$ minimizes $\left\|\beta^p - \hat{\beta}\right\|_{2,A}$. □

Note that it is also possible to only consider the norm of a part of the vector $\beta - \hat{\beta}$ by taking only some entries of each vector, thanks to the next Corollary:

Corollary 5.3.1. *Consider the system* (5.1) *and the Theorem 5.3.3. Let U be a rectangular orthogonal matrix such that each row of U is a unit vector with only one non-zero entry such that $\beta_r = U\beta$ is a reduced vector with only some entries of the vector β. If the additional conditions $S'\beta = m$ (for β^p and $\hat{\beta}^p_c$) or equivalently $S'_r\beta_r = m$ with $\beta_r = U\beta$, and $S' = S'_rU$ are such that:*

$$\forall x \in U\,\mathrm{Ker}(X'),\ \forall s \in \mathrm{Ker}(S'_r) : \langle x, s\rangle_{A_r} = x'A_rs = 0$$

for A_r symmetric positive definite or in other words, if $\mathrm{Ker}(S'_r)$ is the orthogonal complement of $U\,\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_{A_r}$. Then:

$$\hat{\beta}^p_{c,r} = \underset{\substack{\hat{\beta}_r = U\hat{\beta} \\ XV^\dagger X'\hat{\beta} = XV^\dagger Y}}{\arg\min} \left\|\beta^p_r - \hat{\beta}_r\right\|_{2,A_r} = \underset{\substack{\hat{\beta}_r = U\hat{\beta} \\ XV^\dagger X'\hat{\beta} = XV^\dagger Y}}{\arg\min} \sqrt{(\beta^p_r - \hat{\beta}_r)'A_r(\beta^p_r - \hat{\beta}_r)}$$

for the induced norm $\|\cdot\|_{A_r} = \sqrt{\langle\cdot,\cdot\rangle_{A_r}}$

Proof. Let us remind that the set of β's is the set such that $\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon}$ and the set of $\hat{\beta}$'s is the set such that $XV^\dagger Y = XV^\dagger X'\hat{\beta}$. It means that any reduced vector β_r and $\hat{\beta}_r$ can be written as:

$$\begin{aligned}
&\beta_r = U\beta = U\beta_0 + Uv_{\tilde{X}'} && \text{for } \tilde{Y} = \tilde{X}'\beta_0 + \tilde{\epsilon} \\
&\hat{\beta}_r = U\hat{\beta} = U\hat{\beta}_0 + Uv_{XV^\dagger X'} && \text{for } XV^\dagger Y = XV^\dagger X'\hat{\beta}_0
\end{aligned}$$

where $v_{\tilde{X}'} \in \mathrm{Ker}(\tilde{X}')$ and $v_{XV^\dagger X'} \in \mathrm{Ker}(XV^\dagger X')$. It holds furthermore that $\mathrm{Ker}(XV^\dagger X') = \mathrm{Ker}(\tilde{X}')$ since $XV^\dagger X' = \tilde{X}\tilde{X}'$ and from (5.11) we know that $\mathrm{Ker}(\tilde{X}\tilde{X}') = \mathrm{Ker}(\tilde{X}')$. Moreover, $\mathrm{Ker}(X') = \mathrm{Ker}(\tilde{X}')$ as $\mathrm{Ker}(X') \in \mathrm{Ker}(\tilde{X}') =$

$\mathrm{Ker}(T'X')$ and $\mathrm{rank}(XT) = rank(XG) = rank(XP) = rank(X) = p'$ with P and G defined by (5.7). The above expression can thus be written as

$$\begin{aligned}
\beta_r = U\beta = U\beta_0 + Uv_{X'} = \beta_{0,r} + w \qquad & \text{for } \tilde{Y} = \tilde{X}'\beta_0 + \tilde{\epsilon} \\
\hat{\beta}_r = U\hat{\beta} = U\hat{\beta}_0 + Uv_{X'} = \hat{\beta}_{0,r} + w \qquad & \text{for } XV^\dagger Y = XV^\dagger X'\hat{\beta}_0
\end{aligned} \tag{5.13}$$

where $v_{X'} \in \mathrm{Ker}(X')$ and $w = Uv_{x'} \in U\,\mathrm{Ker}(X')$.

Let us define $\hat{\beta}^p_{c2,r} \neq \hat{\beta}^p_{c,r}$, any $\hat{\beta}_r = U\hat{\beta}$ different from $\hat{\beta}^p_{c,r}$ such that $XV^\dagger X'\hat{\beta} = XV^\dagger Y$. Thanks to equation (5.13), it can be written as:

$$\begin{aligned}
\beta^p_r &= \beta_{0,r} + w \\
\hat{\beta}^p_{c,r} &= \hat{\beta}_{0,r} + w \\
\hat{\beta}^p_{c2,r} &= \hat{\beta}_{c,r} + \tilde{w}
\end{aligned}$$

with $\hat{\beta}_{0,r}, \beta_{0,r} \in \mathrm{Ker}(S')$ (by taking $\hat{\beta}_{0,r} = U\beta_0$, $\beta_{0,r} = U\beta_0$ and with $S' = S'_r U$) and with $\tilde{w}, w \in U\,\mathrm{Ker}(X')$ (where $w = Uv_{X'} = UB(S'B)^{-1}m$ using Theorem 5.3.3 and $\tilde{w} = U\tilde{v}_{X'}$ with $\tilde{v}_{X'} \in \mathrm{Ker}(X')$).

We can develop the square of the norm for $\hat{\beta}_{d,r}$ since minimizing the square of a norm is equivalent to minimize the norm:

$$\begin{aligned}
\left\|\beta^p_r - \hat{\beta}^p_{c2,r}\right\|^2_{2,A_r} &= \left\|\beta^p_r - \left(\hat{\beta}^p_{c,r} + \tilde{w}\right)\right\|^2_{2,A_r} \\
&= \left\|\beta^p_r - \hat{\beta}^p_{c,r}\right\|^2_{2,A_r} + \|\tilde{w}\|^2_{2,A_r} - 2\left\langle \beta^p_r - \hat{\beta}^p_{c,r}, \tilde{w}\right\rangle_{A_r} \\
&= \left\|\beta^p_r - \hat{\beta}^p_{c,r}\right\|^2_{2,A_r} + \|\tilde{w}\|^2_{2,A_r} - 2\left\langle \beta_{0,r} - \hat{\beta}_{0,r}, \tilde{w}\right\rangle_{A_r} \\
&= \left\|\beta^p_r - \hat{\beta}^p_{c,r}\right\|^2_{2,A_r} + \|\tilde{w}\|^2_{2,A_r} \\
&> \left\|\beta^p_r - \hat{\beta}^p_{c,r}\right\|^2_{2,A_r} \quad \text{for } \tilde{w} = \hat{\beta}^p_{c2,r} - \hat{\beta}^p_{c,r} \neq \mathbf{0}
\end{aligned}$$

In the above computations, we used the fact that $\left\langle \beta_{0,r} - \hat{\beta}_{0,r}, \tilde{w}\right\rangle_{A_r} = 0$ since $\beta_{0,r} - \hat{\beta}_{0,r} \in \mathrm{Ker}(S')$ and $\tilde{w} \in U\,\mathrm{Ker}(X')$ and $\mathrm{Ker}(S')$ is the orthogonal complement of $U\,\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_{A_r}$. It proves that if $\mathrm{Ker}(S')$ is the orthogonal complement of $U\,\mathrm{Ker}(X')$ for the inner product $\langle\cdot,\cdot\rangle_{A_r}$, $\hat{\beta}^p_{c,r}$ minimizes $\left\|\beta^p_r - \hat{\beta}_r\right\|_{2,A_r}$. □

Chapter 6
Applications to Comparison Models

With the Theorems of the previous Chapter, it is possible to solve the problems stated in Section 3.4 and Chapter 4 namely estimate the item weights $\boldsymbol{w}$ and the additional parameters of the model considered $\boldsymbol{\theta}$ such that

$$\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$$

where $\boldsymbol{z} = \begin{bmatrix}(\log \boldsymbol{\theta})' & (\log \boldsymbol{w})'\end{bmatrix}'$ and M' being

- in the BTL case for pairwise comparisons, $M' = B'$
- in the case for pairwise comparisons with ties, $M' = D \begin{bmatrix}1 & 0\\ 0 & B'\end{bmatrix}$
- in the case of triple comparison, $M' = DB'$

where D is a full rank matrix and B' is the incidence matrix of the simple graph $G = (V, E)$ where V represents the items and E is the set of edges linking two items that are compared to one another in a pair or a triplet.

Results from Chapter 5 allow us to justify, evaluate and complete the extended-method presented in Chapter 4 which was to solve in the least squares sense the noisy (weighted) system:

$$(W^\dagger)^{1/2} \log \boldsymbol{R} \approx (W^\dagger)^{1/2} M'\boldsymbol{z} \;\Leftrightarrow\; MW^\dagger \log \boldsymbol{R} = MW^\dagger M'\hat{\boldsymbol{z}}$$

where the weight matrix W (such that $(W^\dagger) = (W^\dagger)^{-1/2}(W^\dagger)^{-1/2}$) equal to:

- I for the least squares method
- the covariance matrix of $\log \boldsymbol{R}$ for the covariance-weighted least squares method (approximated using an iterative method, theoretically by V or its estimator by $\hat{V}$)
- the diagonal matrix with the variance vector of $\log \boldsymbol{R}$ on the diagonal for the variance-weighted least squares method (approximated using an iterative method, theoretically by D_V or its estimator $\hat{D}_V$)

One solution, satisfying $\sum_{i=1}^{N} \log w_i = 0$, was given by:

$$\hat{\boldsymbol{z}} = \left(MW^\dagger M'\right)^\dagger MW^\dagger \log \boldsymbol{R}$$

In order to use results of the previous Chapter, Section 6.1 characterizes the system $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$ for each considered model, i.e., determine the rank of the matrix M', if the matrix V is full rank and if not, check that $\mathrm{Ker}(V) \subseteq \mathrm{Ker}(M)$, and determine the kernel of the matrix M. Section 6.2 presents several error measures relevant for our application, and Section 6.3 details our general method for comparison models and its performance regarding the different error measures.

6.1 Characterization of the model systems

6.1.1 BTL Model

The equations The BTL model does not need any additional parameter, meaning that $\boldsymbol{z} = \log \boldsymbol{w}$. The equations used for this model are given by

$$\log R_{i,j} \approx \log w_i - \log w_j$$

The coefficient matrix *M'* is a matrix of dimension $n_{eq} \times N$ equal to B' which represents the incidence matrix of a connected simple graph, meaning that its rank is $N - 1$. It also holds that $M'\mathbb{1} = 0$.

The error variance *V* has a full rank, as the comparison of two independent pairs of items are independent as well.

6.1.2 Application to pairwise comparison models with ties

For the two models that take ties into account for pairwise comparisons, the equations allowing to recover item weights are such that M' is of rank $\tilde{N} - 1 = N$ (there is one additional parameter θ besides the item weights $\boldsymbol{w}$). In both cases, it indeed holds that

$$M' = D \begin{bmatrix} 1 & 0 \\ 0 & B' \end{bmatrix} = D\tilde{B}'$$

where B' is the incidence matrix of the simple graph $G = (V, E)$ such that the vertices are all the items and E links to items together if they are compared together and D is a full rank matrix of size $n_{eq} \times (n_s + 1)$ (n_s is the total amount of pairs). Using the rank properties, we can use that fact to deduce the rank of M':

$$\begin{aligned} \operatorname{rank}(D) + \operatorname{rank}(\tilde{B}') - (n_s + 1) \leq\ & \operatorname{rank}(M') \leq \min\left\{\operatorname{rank}(D), \operatorname{rank}(\tilde{B}')\right\} \\ N \leq\ & \operatorname{rank}(M') \leq N \end{aligned}$$

since $\operatorname{rank}(\tilde{B}') = N < n_s + 1$. The details about the rank of the matrix D for each model is given in Appendix D.

Rao&Kupper Model

The equations for this model are given by

$$\begin{aligned} \log \frac{h_{i,j}}{1 - h_{i,j}} &\approx \log w_i - \log w_j - \log \theta \\ \log \frac{h_{j,i}}{1 - h_{j,i}} &\approx \log w_j - \log w_i - \log \theta \end{aligned}$$

The coefficient matrix *M'* is given by

$$M' = \begin{bmatrix} -\mathbb{1} & B' \\ -\mathbb{1} & -B' \end{bmatrix} = \begin{bmatrix} -\mathbb{1}_{n_s} & I_{n_s} \\ -\mathbb{1}_{n_s} & -I_{n_s} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & B' \end{bmatrix}$$

which has a rank $N = \tilde{N} - 1$ as explained above (See Appendix D for further details). With the above matrix, we observe that $M' \begin{bmatrix} 0 \\ \mathbb{1} \end{bmatrix} = 0$.

The error variance *V* has a full rank since all the equations are independent.

Davidson Model

The equations for this model are given by

$$\begin{aligned} \log \frac{h_{i,j}}{h_{j,i}} &\approx \log w_i - \log w_j \\ \log \frac{h_{i=j}}{h_{i,j}} &\approx \log \theta + \frac{1}{2} \log w_j - \frac{1}{2} \log w_i \\ \log \frac{h_{i=j}}{h_{j,i}} &\approx \log \theta + \frac{1}{2} \log w_i - \frac{1}{2} \log w_j \end{aligned}$$

The coefficient matrix *M'* is given by

$$M' = \begin{bmatrix} 0 & B' \\ \mathbb{1} & -\frac{1}{2}B' \\ \mathbb{1} & \frac{1}{2}B' \end{bmatrix} = \begin{bmatrix} 0_{n_s} & I_{n_s} \\ \mathbb{1}_{n_s} & -\frac{1}{2}I_{n_s} \\ \mathbb{1}_{n_s} & \frac{1}{2}I_{n_s} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & B' \end{bmatrix}$$

The matrix M' has a rank $N = \tilde{N} - 1$ as explained previously (See Appendix D for further details). With the above matrix, we observe that $M' \begin{bmatrix} 0 \\ \mathbb{1} \end{bmatrix} = 0$.

The error variance *V* is in this case a singular matrix. One can indeed observe that among the three equations given for each pair of items, only two of them are independent. One has thus that V of dimension $n_{eq} \times n_{eq}$ (with $n_{eq} = 3n_s$) has a rank of $2n_s$. As explained above, we need in this case to have $\text{Ker}(V) \subseteq \text{Ker}(M)$ to be able to find the BLUE of $\boldsymbol{z}$. To do so, we will find a basis F of $\text{Ker}(V)$ leading to $F'V = 0$. This matrix will of dimensions $n_{eq} \times n_s$.

Suppose the equations are ranked in the following order:

$$\left\{ \frac{\boldsymbol{h_{i,j}}}{\boldsymbol{h_{j,i}}}, \frac{\boldsymbol{h_{i=j}}}{\boldsymbol{h_{i,j}}}, \frac{\boldsymbol{h_{i=j}}}{\boldsymbol{h_{j,i}}} \right\} \qquad \text{where } \frac{\boldsymbol{h_{i,j}}}{\boldsymbol{h_{j,i}}} = \left(\frac{h_{1,2}}{h_{2,1}}, \frac{h_{1,3}}{h_{3,1}}, \cdots, \frac{h_{N-1,N}}{h_{N,N-1}} \right)$$

A matrix F such that $F'V = 0$, i.e. basis of $\text{Ker}(V)$, can be

$$F = I \bigotimes \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} = \begin{pmatrix} I \\ I \\ -I \end{pmatrix} \qquad \text{because} \quad \log \left(\frac{h_{i,j}}{h_{j,i}} \right) + \log \left(\frac{h_{i=j}}{h_{i,j}} \right) = \log \left(\frac{h_{i=j}}{h_{j,i}} \right)$$

where I is the identity matrix of dimension n_s. Furthermore, we have that $\text{Ker}(V) \subseteq \text{Ker}(M)$ as $F'M' = 0$ with the matrix M' given above.

6.1.3 Application to Triple comparison Model

The two triple comparison models considered also have equations with a coefficient matrix M' of rank $\tilde{N} - 1 = N - 1$. It holds indeed in both cases that

$$M' = DB'$$

where B' is the incidence matrix of the simple graph with the items as vertices and that links two vertices if they belong to at least one triplet of items being compared to one another and where D is a full rank matrix of size $n_{eq} \times n_{edge}$ (where n_{edge} represents the number of edges composing the graph). Since the rank of B' is equal to $N - 1$ as we only consider connected graphs ($n_{edge} > N - 1$), it holds that

$$\begin{aligned} \text{rank}(D) + \text{rank}(B') - n_{edge} \leq \ & \text{rank}(M') \leq \min\Big\{\text{rank}(D), \text{rank}(B')\Big\} \\ N - 1 \leq \ & \text{rank}(M') \leq N - 1 \end{aligned}$$

The details about the matrix D for each model is given in Appendix D.

Plackett-Luce Model

The equations are given by

$$\log\left(\frac{h_{i,j}}{h_{j,i}}\right) \approx \log(w_i) - \log(w_j)$$

for all pairs $(i, j) \in \{(i, j), (i, k), (j, k)\}$ and all triplet of items (i, j, k).

The coefficient matrix *M'* is composed of rows equal to $e_i - e_j$ where e_i is the null vector with entry i equal to 1 which can be written as $M' = DB'$ with D of full rank. It holds that M' is of rank $N - 1$ as explained previously (see Appendix D for the matrix D) and that $M'\mathbb{1} = 0$.

The error variance *V* has a full rank. As explained in Chapter 4, the equations are all independent.

Mallows-Bradley-Terry Model

The equations are given by

$$\log\left(\frac{h_\pi}{h_{\pi'}}\right) \approx (\pi'^{-1} - \pi^{-1})(i)\log(w_i) + (\pi'^{-1} - \pi^{-1})(j)\log(w_j) + (\pi'^{-1} - \pi^{-1})(k)\log(w_k)$$

for all rankings $\pi, \pi' \in \{(i, j, k), (i, k, j), (j, i, k), (j, k, i), (k, i, j), (k, j, i)\}$ and all triplet of items (i, j, k).

The coefficient matrix *M'* of rank $N - 1$ satisfy $M'\mathbb{1} = 0$ since each row of M' sums to 0. The coefficients can indeed be seen as the commutation of indices in the two rankings compared.

The error variance $\boldsymbol{V}$ of dimension $n_{eq} \times n_{eq}$ (with $n_{eq} = 15n_s$, n_s being the total number of triplets) has a rank of $5n_s$. Regarding one triplet of items, there are 5 independent equations on the 15 ones. In order to find the BLUE estimator of $\boldsymbol{z}$ in this case, we need to find a basis F of $\text{Ker}(V)$, i.e. $F'V = 0$. Since V is block diagonal as explained in Chapter 4, F which has dimension $n_{eq} \times 10n_s$ can be a block diagonal as well

$$F = \tilde{F} \bigotimes I = \begin{bmatrix} \tilde{F} & & \\ & \ddots & \\ & & \tilde{F} \end{bmatrix} \qquad \text{with } \tilde{F} \text{ of dimension } 15 \times 10$$

The matrix $\tilde{F}$ needs to be such that $\tilde{F}'\tilde{V} = 0$ where $\tilde{V}$ is one block of the variance matrix corresponding to one triplet of items i, j, k. Suppose the equations are ranked in the following order for each triplet of items i, j, k:

$$\left\{ \frac{h_{i,k,j}}{h_{i,j,k}}, \frac{h_{j,i,k}}{h_{i,j,k}}, \frac{h_{j,k,i}}{h_{i,j,k}}, \frac{h_{k,i,j}}{h_{i,j,k}}, \frac{h_{k,j,i}}{h_{i,j,k}}, \frac{h_{j,i,k}}{h_{i,k,j}}, \frac{h_{j,k,i}}{h_{i,k,j}}, \frac{h_{k,i,j}}{h_{i,k,j}}, \right.$$
$$\left. \frac{h_{k,j,i}}{h_{i,k,j}}, \frac{h_{j,k,i}}{h_{j,i,k}}, \frac{h_{k,i,j}}{h_{j,i,k}}, \frac{h_{k,j,i}}{h_{j,i,k}}, \frac{h_{k,i,j}}{h_{j,k,i}}, \frac{h_{k,j,i}}{h_{j,k,i}}, \frac{h_{k,j,i}}{h_{k,i,j}} \right\}$$

A basis $\tilde{F}$ of $\text{Ker}(\tilde{V})$ is given by

$$\tilde{F} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & -1 & -1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

because

$$\log\left(\frac{h_\pi}{h_{i,j,k}}\right) - \log\left(\frac{h_{\pi'}}{h_{i,j,k}}\right) = \log\left(\frac{h_\pi}{h_{\pi'}}\right)$$

for $\pi, \pi' \in \{(i,k,j), (j,i,k), (j,k,i), (k,i,j), (k,j,i)\}$.

We observe with the latter relation that $F'M' = 0$ since

$$\log\left(\frac{h_\pi}{h_{i,j,k}}\right) \approx (1-\pi^{-1}(i))\log(w_i) + (2-\pi^{-1}(j))\log(w_j) + (3-\pi^{-1}(k))\log(w_k)$$
$$\log\left(\frac{h_{\pi'}}{h_{i,j,k}}\right) \approx (1-\pi'^{-1}(i))\log(w_i) + (2-\pi'^{-1}(j))\log(w_j) + (3-\pi'^{-1}(k))\log(w_k)$$
$$\log\left(\frac{h_\pi}{h_{\pi'}}\right) \approx (\pi'^{-1}-\pi^{-1})(i)\log(w_i) + (\pi'^{-1}-\pi^{-1})(j)\log(w_j) + (\pi'^{-1}-\pi^{-1})(k)\log(w_k)$$

The coefficients of $\log w_*$ from the right terms also satisfy the relation

$$\log\left(\frac{h_\pi}{h_{i,j,k}}\right) - \log\left(\frac{h_{\pi'}}{h_{i,j,k}}\right) = \log\left(\frac{h_\pi}{h_{\pi'}}\right)$$

6.2 Error measures

Different error measures can be considered when looking for item scores close to the real ones. As mentioned previously, we know that the vectors $\boldsymbol{z}^p$ belong to a line as their estimators $\hat{\boldsymbol{z}}^p$. Indeed, the weights $\boldsymbol{w}$ that we want to recover are unobservable. We observe in the different models that multiplying the parameters $\boldsymbol{w}$ by a scalar α does not modify the probabilities.

This leads to a first error measure, the angle between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$, which will be independent of the additional constraint added to the problem which renders the parameters observable. One could also be interested in either minimizing the error $\boldsymbol{z} - \hat{\boldsymbol{z}}$, $\log \boldsymbol{w} - \log \hat{\boldsymbol{w}}$ or $\boldsymbol{w} - \hat{\boldsymbol{w}}$. Those error measures are detailed below.

Angle between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$ The first error measure considered is the smallest angle between the two vectors or more precisely its sine to have a measure decreasing when the error decreases. This angle is an error measure which is invariant to the chosen constraint for the estimator and the scale of $\boldsymbol{w}$ which allows us to look for an estimation $\hat{\boldsymbol{w}}$ close in a scale-invariant sense to the true weights $\boldsymbol{w}$. The invariant property can be proven thanks to the definition of the cosine, since angles $\in [0°, 180°]$ with the same cosines are equal:

$$\cos(\alpha\boldsymbol{w}, \hat{\alpha}\hat{\boldsymbol{w}}) = \frac{\langle \alpha\boldsymbol{w}, \hat{\alpha}\hat{\boldsymbol{w}}\rangle}{\|\alpha\boldsymbol{w}\|\|\hat{\alpha}\hat{\boldsymbol{w}}\|} = \frac{\alpha\hat{\alpha}\langle \boldsymbol{w}, \hat{\boldsymbol{w}}\rangle}{\alpha\hat{\alpha}\|\boldsymbol{w}\|\|\hat{\boldsymbol{w}}\|} = \frac{\langle \boldsymbol{w}, \hat{\boldsymbol{w}}\rangle}{\|\boldsymbol{w}\|\|\hat{\boldsymbol{w}}\|} = \cos(\boldsymbol{w}, \hat{\boldsymbol{w}}) \tag{6.1}$$

We detailed a method to find an estimator of a particular vector $\log \boldsymbol{w}$ satisfying one additional condition and discussed the condition to choose but this measure allows us to evaluate the performances of the general method, i.e. the least squares approach, regardless of the chosen constraint.

Norm of $\Delta\hat{\boldsymbol{z}}$ We are also interested into the impact of the chosen constraint for the estimator and since we solve a system of parameters $\boldsymbol{z}$, another measure of the error would be the norm of this vector, we can also consider any weighted norm $\|\boldsymbol{e}\|_A = \sqrt{\boldsymbol{e}'A\boldsymbol{e}}$.

Norm of a part of $\Delta\hat{\boldsymbol{z}}$ ($\Delta \log \hat{\boldsymbol{w}}$) We could also look at the error on a part of the vector $\Delta\hat{\boldsymbol{z}} = \boldsymbol{z} - \hat{\boldsymbol{z}}$, such as only $\Delta \log \hat{\boldsymbol{w}} = \log \boldsymbol{w} - \log \hat{\boldsymbol{w}}$ which refers to the item scores. Again, any 2−norm can be considered.

Norm of $\Delta\hat{\boldsymbol{w}}$ Finally, one could also be interested in minimizing the norm of the error on $\Delta\boldsymbol{w} = \boldsymbol{w} - \hat{\boldsymbol{w}}$ but then it would be more accurate to minimize it for the relative error $\Delta\boldsymbol{w}/\|\boldsymbol{w}\|$ in order to have the same minimum for any scaled $\boldsymbol{w}$.

Notice that, for those last measures, the minimum value of the error will not be impacted by the scale of $\boldsymbol{w}$. Denote $\hat{\boldsymbol{w}}_{min}$ the estimator minimizing the chosen error measure with a specific $\boldsymbol{w}$ and with a new scaled version $\alpha\boldsymbol{w}$ of $\boldsymbol{w}$, the minimum becomes:

$$\begin{aligned}
\min_{\hat{\boldsymbol{w}}} \|\log \boldsymbol{w} - \log \hat{\boldsymbol{w}}\| &= \|\log \boldsymbol{w} - \log \hat{\boldsymbol{w}}_{min}\| \\
\min_{\hat{\boldsymbol{w}}} \|\log \alpha\boldsymbol{w} - \log \hat{\boldsymbol{w}}\| &= \min_{\hat{\boldsymbol{w}}} \left\| \log \alpha\boldsymbol{w} - \log \frac{\alpha\hat{\boldsymbol{w}}}{\alpha} \right\| = \min_{\hat{\boldsymbol{w}}} \left\| \log \boldsymbol{w} - \log \frac{\hat{\boldsymbol{w}}}{\alpha} \right\| \\
&= \left\| \log \boldsymbol{w} - \log \frac{\hat{\boldsymbol{w}}_{new,min}}{\alpha} \right\| = \|\log \boldsymbol{w} - \log \hat{\boldsymbol{w}}_{min}\|
\end{aligned}$$

$$\begin{aligned}
\min_{\hat{\boldsymbol{w}}} \frac{\|\boldsymbol{w} - \hat{\boldsymbol{w}}\|}{\|\boldsymbol{w}\|} &= \frac{\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{min}\|}{\|\boldsymbol{w}\|} \\
\min_{\hat{\boldsymbol{w}}} \frac{\|\alpha\boldsymbol{w} - \hat{\boldsymbol{w}}\|}{\|\alpha\boldsymbol{w}\|} &= \min_{\hat{\boldsymbol{w}}} \frac{\left\| \alpha\boldsymbol{w} - \frac{\alpha\hat{\boldsymbol{w}}}{\alpha} \right\|}{\|\alpha\boldsymbol{w}\|} = \min_{\hat{\boldsymbol{w}}} \frac{\left\| \boldsymbol{w} - \frac{\hat{\boldsymbol{w}}}{\alpha} \right\|}{\|\boldsymbol{w}\|} \\
&= \frac{\left\| \boldsymbol{w} - \frac{\hat{\boldsymbol{w}}_{new,min}}{\alpha} \right\|}{\|\boldsymbol{w}\|} = \frac{\|\boldsymbol{w} - \hat{\boldsymbol{w}}_{min}\|}{\|\boldsymbol{w}\|}
\end{aligned}$$

by taking $\hat{\boldsymbol{w}}_{new,min} = \alpha\hat{\boldsymbol{w}}_{min}$.

The minimum value of those error measures is invariant to the scale of $\boldsymbol{w}$ and depends on the scale choice of the estimator $\hat{\boldsymbol{w}}$.

6.3 General method

In all the considered models, the system to approximate, $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$, has a coefficient matrix M' of dimension $n_{eq} \times \tilde{N}$ and of rank $\tilde{N} - 1$, as shown in Section 6.1. This means that $\boldsymbol{z}$ satisfying the above problem belongs thus to a line (affine set associated to a vector space of dimension one). The matrix M' is such that $v = \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix}'$ belongs the kernel of M' and forms therefore a basis of the vector subspace associated to this affine subspace.

As stated in Chapter 5, having a non-full rank matrix implies that there is no definition of an unbiased estimator of $\boldsymbol{z}$ since $\boldsymbol{z}$ is not unique. However, one can uniquely identify one particular $\boldsymbol{z}^p$ by adding one condition (since they belong to a line) that the parameters must satisfy. In that way, we can find the BLUE estimator of the specific $\boldsymbol{z}^p$ satisfying the additional condition, in the constrained subspace thanks to

Theorem 5.3.3. Based on this result, we can also deduce the linear unbiased estimator which minimizes any specific 2−norm thanks to Theorem 5.3.4. We also go further by finding some relevant information about estimators minimizing other error measures, more specific to our application, detailed in the previous Section.

6.3.1 BLUE estimator of $\boldsymbol{z}^p$ in a constrained subspace C

As explained previously, we can focus on any particular parameters vector $\boldsymbol{z}^p$, uniquely identified by an additional condition. Let this condition be of the form $S'\boldsymbol{z} = m$ such that $\boldsymbol{z}^p \in C = \{x \in \mathbb{R}^{\tilde{N}} | S'x = m\}$ with S of dimension $1 \times \tilde{N}$ and Q' be an orthonormal basis of $\mathrm{Ker}(S')$.

Using this additional constraint, one can find the BLUE estimator of $\boldsymbol{z}^p$ satisfying the same additional condition by applying the result of Theorem 5.3.3:

$$\hat{\boldsymbol{z}}_c^p = \left(Q'QMV^\dagger M'Q'Q\right)^\dagger MV^\dagger \log \boldsymbol{R} + \frac{m}{\sum_{i=1}^N S_{i+\tilde{N}-N}} \begin{bmatrix} 0_{\tilde{N}-N} \\ \mathbb{1}_N \end{bmatrix} \tag{6.2}$$

where V is the covariance matrix of the vector $\log \boldsymbol{R}$. It means that among all the vectors satisfying $S'\boldsymbol{z} = m$, the vector $\hat{\boldsymbol{z}}_c^p$ is the best linear unbiased estimator of $\boldsymbol{z}^p$, i.e. it minimizes the variance of $\Delta\hat{\boldsymbol{z}}_c^p = \hat{\boldsymbol{z}}_c^p - \boldsymbol{z}^p$.

The expectation of $\|\Delta\hat{\boldsymbol{z}}_c^p\|^2$ is also minimum in the constrained subset of estimators since:

$$\mathbb{E}\left[\|\Delta\hat{\boldsymbol{z}}_c^p\|_2^2\right] = \mathrm{Tr}\left(\mathbb{E}\left[\Delta\hat{\boldsymbol{z}}_c^p \ (\Delta\hat{\boldsymbol{z}}_c^p)'\right]\right) = \mathrm{Tr}\left(\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right]\right) \tag{6.3}$$

For any other linear unbiased estimator $\tilde{\boldsymbol{z}}_c^p$ in the constrained space, we have by the properties of the trace [1]:

$$\begin{aligned} \mathbb{E}\left[\|\Delta\tilde{\boldsymbol{z}}_c^p\|_2^2\right] &= \mathrm{Tr}\left(\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right]\right) \\ &= \mathrm{Tr}\left(\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right]\right) + \mathrm{Tr}\left(\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] - \mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right]\right) \\ &\leq \mathrm{Tr}\left(\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right]\right) = \mathbb{E}\left[\|\Delta\hat{\boldsymbol{z}}_c^p\|_2^2\right] \end{aligned} \tag{6.4}$$

since $\mathbb{V}\left[\tilde{\boldsymbol{z}}_c^p - \boldsymbol{z}^p\right] \succeq \mathbb{V}\left[\hat{\boldsymbol{z}}_c^p - \boldsymbol{z}^p\right]$ by the definition of BLUE.

The expected value of this error can be computed as well. Define $\log \boldsymbol{\rho} = M'\boldsymbol{z}^p$ where $\boldsymbol{\rho}$ contains the true ratio between item scores, for example in the BTL model $\rho_{i,j} = w_i/w_j$. Then

$$\Delta\hat{\boldsymbol{z}}_c^p = \hat{\boldsymbol{z}}_c^p - \boldsymbol{z}^p = \left(Q'QMV^\dagger M'Q'Q\right)^\dagger MV^\dagger(\log \boldsymbol{R} - \log \boldsymbol{\rho})$$

where $\log \boldsymbol{R} - \log \boldsymbol{\rho} = (M'\boldsymbol{z}^p + \boldsymbol{\epsilon}) - (M'\boldsymbol{z}^p) = \boldsymbol{\epsilon}$ and it holds that

$$\begin{aligned} \mathbb{E}\left[\Delta\hat{\boldsymbol{z}}_c^p \ (\Delta\hat{\boldsymbol{z}}_c^p)'\right] &= \left(Q'QMV^\dagger M'Q'Q\right)^\dagger MV^\dagger \mathbb{E}\left[\tilde{\boldsymbol{\epsilon}} \ \tilde{\boldsymbol{\epsilon}}'\right] V^\dagger M' \left(Q'QMV^\dagger M'Q'Q\right)^\dagger \\ &= Q'\left(QMV^\dagger M'Q'\right)^\dagger QMV^\dagger VV^\dagger M'Q'\left(QMV^\dagger M'Q'\right)^\dagger Q \\ &= \left(Q'QMV^\dagger M'Q'Q\right)^\dagger \end{aligned}$$

[1] $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$

where the last expressions are obtained using the property of the Moore–Penrose inverse, namely $A^\dagger A A^\dagger = A^\dagger$ and the fact that $Q' = Q^\dagger$[2]. The expected error (6.3) is thus obtained by taking the trace of this matrix.

It holds in fact that the expectation of any 2−norm of the error $\Delta\hat{\boldsymbol{z}}_c^p$, defined by $\|\boldsymbol{z}^p - \hat{\boldsymbol{z}}_c^p\|_{2,A} = \sqrt{(\Delta\hat{\boldsymbol{z}}_c^p)' A \Delta\hat{\boldsymbol{z}}_c^p}$ for a matrix A symmetric positive definite, is also minimized in the constrained subspace:

$$\begin{aligned}\mathbb{E}\left[\,\|\Delta\hat{\boldsymbol{z}}_c^p\|_{2,A}^2\right] = \mathbb{E}\left[\,(\Delta\hat{\boldsymbol{z}}_c^p)' A \Delta\hat{\boldsymbol{z}}_c^p\right] &= \mathbb{E}\left[\,\left\|D'\Delta\hat{\boldsymbol{z}}_c^p\right\|_2^2\right] \qquad \text{with } A = DD' \\ &= \mathrm{Tr}\left(D'\;\mathbb{E}\left[\Delta\hat{\boldsymbol{z}}_c^p\;(\Delta\hat{\boldsymbol{z}}_c^p)'\right] D\right) = \mathrm{Tr}\left(D'\;\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] D\right)\end{aligned}$$

From the definition of BLUE, we have that for any linear unbiased estimator $\tilde{\boldsymbol{z}}_{\boldsymbol{c}}^{\boldsymbol{p}}$ of $\boldsymbol{z}^{\boldsymbol{p}}$ in the constrained subspace:

$$\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] \succeq \mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] \Leftrightarrow x'\,\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] x \geq x'\,\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] x \quad \forall x$$

It also holds that:

$$D'\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] D \succeq D'\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] D \Leftrightarrow x'D'\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] Dx \geq x'D'\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] Dx \quad \forall x$$

since $y'\,\mathbb{V}\left[\Delta\tilde{\boldsymbol{z}}_c^p\right] y \geq y'\,\mathbb{V}\left[\Delta\hat{\boldsymbol{z}}_c^p\right] y$ with $y = Dx$. In the same manner than (6.4), this norm of the error $\Delta\hat{\boldsymbol{z}}_c^p$ is minimum among all linear unbiased constrained estimators, i.e.

$$\mathbb{E}\left[\,\|\Delta\hat{\boldsymbol{z}}_c^p\|_{2,A}^2\right] \leq \mathbb{E}\left[\,\|\Delta\tilde{\boldsymbol{z}}_c^p\|_{2,A}^2\right] \tag{6.5}$$

6.3.2 Minimum $\|\Delta\hat{\boldsymbol{z}}^p\|$ error estimator

If there exists a BLUE estimator of a particular vector β, this estimator would be the one minimizing any A−weighted 2−norm of the error among all the linear unbiased estimators. This has been proven in the previous Section by the relation (6.5) in the constrained space for the vector $\boldsymbol{z}^{\boldsymbol{p}}$. In other words, if $\boldsymbol{z}^p$ has a best linear unbiased estimator, it would minimize $\|\boldsymbol{z}^{\boldsymbol{p}} - \hat{\boldsymbol{z}}^{\boldsymbol{p}}\|_{2,A}$ for any matrix A positive definite.

As mentioned previously in Chapter 5, for a particular $\boldsymbol{z}^p$, the set formed by the BLUE estimator in each possible constrained subspace, defined by $S'\boldsymbol{z} = m$, is a line $MV^\dagger \log \boldsymbol{R} = MV^\dagger M'\hat{\boldsymbol{z}}$, and this line is independent of the chosen $\boldsymbol{z}^p$ and parallel to the line defining $\boldsymbol{z}$. The union of all the constrained subspaces identifying $\boldsymbol{z}^p$ forms the whole space except the vectors $\boldsymbol{z} \neq \boldsymbol{z}^p$ from the line of $\boldsymbol{z}$, i.e. $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$. Since no vector $\boldsymbol{z} \neq \boldsymbol{z}^p$ of the line can be an unbiased estimator of $\boldsymbol{z}^p$, it means that all the linear unbiased estimators of $\boldsymbol{z}^p$ are contained in this union of subspaces.

From (6.5) we know that in a constrained subspace, the BLUE of $\boldsymbol{z}^{\boldsymbol{p}}$ minimizes any A−weighted 2−norm. Since the union of all the constrained subspaces contains all the linear unbiased estimators of $\boldsymbol{z}^{\boldsymbol{p}}$, it holds that the estimator minimizing a given weighted 2−norm belongs to that unconstrained space. More precisely, we know that

[2]The property of the Moore-Penrose pseudo-inverse, namely $(AB)^\dagger = B^\dagger A^\dagger$ when A has orthonormal columns ($A'A = A^\dagger A = I$) or when B has orthonormal rows ($BB' = BB^\dagger = I$) is used with the matrix Q which has orthonormal rows.

$\hat{\boldsymbol{z}}^{\boldsymbol{p}}$ minimizing $\|\boldsymbol{z}^p - \hat{\boldsymbol{z}}^{\boldsymbol{p}}\|_{2,A}$ belongs to the line defining the best linear unbiased constrained estimators of $\boldsymbol{z}^{\boldsymbol{p}}$, $MV^\dagger \log \boldsymbol{R} = MV^\dagger M'\hat{\boldsymbol{z}}$ since for each constrained subspace, the estimator minimizing that $A-$weighted 2$-$norm belongs to the line.

By Theorem 5.3.4, we know furthermore that among the best linear unbiased constrained estimator, a unique estimator strictly minimizes a $A-$weighted 2$-$norm. This means that for different estimators found by imposing different constraint on $\boldsymbol{z}^{\boldsymbol{p}}$, two different $A-$weighted 2$-$norm will be strictly minimized. In other words, there exists no linear unbiased estimator of $\boldsymbol{z}^{\boldsymbol{p}}$ that minimizes any $A-$weighted 2$-$norm in the unconstrained space, meaning that a BLUE estimator cannot exist. Nevertheless, the Theorem allows us thus to find the linear unbiased estimator (without constraints to satisfy) minimizing a certain 2$-$norm of the error, i.e., $\sqrt{(\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p)'A(\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p)}$ for a matrix A symmetric positive definite.

In the purpose to find the estimator $\hat{\boldsymbol{z}}^{\boldsymbol{p}}$ of $\boldsymbol{z}^{\boldsymbol{p}}$ minimizing a given $A-$weighted 2$-$norm, Theorem 5.3.4 will be used. It indeed tells us that the estimator satisfying the constraint $S'\boldsymbol{z} = m$ such that

$$\forall x \in \mathrm{Ker}(M'), \forall s \in \mathrm{Ker}(S') : \; x'As = 0$$

will be the one minimizing $\|\boldsymbol{z}^{\boldsymbol{p}} - \hat{\boldsymbol{z}}^{\boldsymbol{p}}\|_{2,A}$ for $\boldsymbol{z}^{\boldsymbol{p}}$ satisfying the same constraint. As previously mentioned, the vector $v = \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix}'$ forms a basis of $\mathrm{Ker}(M')$. The above condition becomes thus to have S' satisfying $v'As = 0$ for all $s \in \mathrm{Ker}(S')$. Note that $\boldsymbol{z}^p = \boldsymbol{z}_0 + \alpha v$ such that $S'\boldsymbol{z}_0 = 0$, $S'(\alpha v) = m$ with $\alpha = (S'v)^{-1}m$ and thus $\boldsymbol{z}_0$ forms a basis of $\mathrm{Ker}(S')$. The condition is therefore of the form:

$$v'As = \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} A\boldsymbol{z}_0 = 0 \implies S' = \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} A$$

Among all the linear unbiased estimators of $\boldsymbol{z}^p$, identified by $\begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} A\boldsymbol{z} = m$, the estimator minimizing the $A-$weighted 2$-$norm with $\boldsymbol{z}^p$, i.e. $\sqrt{(\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p)'A(\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p)}$, is

$$\hat{\boldsymbol{z}}^p = \left(Q'QMV^\dagger M'Q'Q\right)^\dagger MV^\dagger \log \boldsymbol{R} + \frac{m}{\begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} A \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix}'} \begin{bmatrix} 0_{\tilde{N}-N} \\ \mathbb{1}_N \end{bmatrix}$$

with Q as before, an orthonormal basis of $\mathrm{Ker}\left(\begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} A\right)$. This estimator corresponds in fact to the BLUE estimator of $\boldsymbol{z}^{\boldsymbol{p}}$ in the constrained subspace $C = \left\{ x \,\middle|\, \begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} Ax = m \right\}$.

In the case of the simplest 2$-$norm of the error vector, i.e. $\|\boldsymbol{z}^{\boldsymbol{p}} - \hat{\boldsymbol{z}}^{\boldsymbol{p}}\|_{2,A}$ with $A = I$, the constraint becomes

$$\begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} \boldsymbol{z} = \sum_{i=1}^{N} \log w_i = m$$

Note that for all the different models, the rows of M' are such that the last N entries, the entries associated to item weights, sum to zero. This means that $S'M =$

$\begin{bmatrix} 0'_{\tilde{N}-N} & \mathbb{1}'_N \end{bmatrix} M = 0$ and thus $Q'QM = M$. The estimator of $\boldsymbol{z}^p$, such that $\sum_{i=1}^N \log w_i = m$ in this particular case is thus given by :

$$\hat{\boldsymbol{z}}^p = \left(MV^\dagger M'\right)^\dagger MV^\dagger \log \boldsymbol{R} + \frac{m}{N}\begin{bmatrix} 0_{\tilde{N}-N} \\ \mathbb{1}_N \end{bmatrix} \tag{6.6}$$

And the expectation of the error (6.3) becomes $\mathbb{E}\left[\left\|\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p\right\|_2^2\right] = \mathrm{Tr}\left(\left(MV^\dagger M'\right)^\dagger\right)$.

This solution is thus the linear unbiased estimator that minimizes the expectation of $\left\|\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p\right\|_2$, and the best linear unbiased estimator of $\boldsymbol{z}^p$ in the constrained subspace $S'\boldsymbol{z} = m$. Figure A.4 in Appendix A shows the empirical mean and expectation of $\left\|\boldsymbol{z}^p - \hat{\boldsymbol{z}}^p\right\|_2$ for this estimator and another estimator on the line with our method using the empirical estimated approximation of the variance $\hat{V}$.

This solution, for $m = 0$, corresponds to the solution presented in Section 3.4 and Chapter 4 when solving the problem in the least squares sense with a weight matrix corresponding to the pseudo-inverse of the covariance of $\log \boldsymbol{R}$. Remember that for the BTL extensions models, the weight matrix used in the weighted least squares problem could be taken as the covariance matrix of $\log \boldsymbol{R}$ or as the diagonal matrix with the variance vector of $\log \boldsymbol{R}$ on the diagonal. The above solution corresponds to the one of the weighted least squares problem with the covariance as weight matrix. This means that the least squares solution and the weighted one with the diagonal matrix of the variance, except for BTL since the covariance matrix is diagonal, lead to non-optimal solutions, i.e., with a higher expected error norm and variance[3].

Corollary 5.3.1 also allows us to find the linear unbiased estimator minimizing a specific $2-$norm of a part of the vector error, i.e., the error on some entries of this vector $\boldsymbol{z}$.

6.3.3 Theoretical minimum error $\|\Delta\hat{\boldsymbol{w}}^p\| / \|\boldsymbol{w}^p\|$ estimator

The previous results aim to compute the norm of the error on $\Delta\boldsymbol{z}^p$ and minimizes it in the constrained space or not. One could also be interested in the relative error on $\Delta\boldsymbol{w}^p = \boldsymbol{w}^p - \hat{\boldsymbol{w}}^p$ since $\boldsymbol{w}^p$ is initially the parameters we are interested in, even if $\boldsymbol{z}^p = \begin{bmatrix} (\log\boldsymbol{\theta})' & (\log \boldsymbol{w}^p)' \end{bmatrix}'$ is another representation of those parameters.

Using a Taylor first order expansion on $\Delta\boldsymbol{w}^p / \|\boldsymbol{w}^p\|$ we obtain the following approximation:

$$\frac{\|\Delta\boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|} \approx \left\| \mathrm{diag}\left(\frac{\boldsymbol{w}^p}{\|\boldsymbol{w}^p\|}\right) \Delta \log \boldsymbol{w}^p \right\| = \left\| \mathrm{diag}\left(\begin{bmatrix} 0'_{\tilde{N}-N} & \frac{\boldsymbol{w}^p}{\|\boldsymbol{w}^p\|} \end{bmatrix}\right) \Delta\boldsymbol{z}^p \right\| = \|\Delta\boldsymbol{z}_r\|_{D_w^2}$$

with $D_w = \mathrm{diag}\left(\frac{\boldsymbol{w}^p}{\|\boldsymbol{w}^p\|}\right)$, $\Delta\boldsymbol{z}_r = \Delta \log \boldsymbol{w}^p$.

To minimize this error approximation, we would need to compute the linear unbiased estimator of a part of $\boldsymbol{z}$ minimizing the D_w^2-weighted $2-$norm. Note that this approach

[3]Estimator with higher variance means here that the variance of its error, $\tilde{V}$, is such that $\tilde{V} - V \succeq 0$ with V the variance of the other estimator error.

can only be used theoretically when the true weights are known.

Remember from the previous Section that in our cases, if we want to minimize the $A-$weigthed 2$-$norm, the constraint is of the form $\begin{bmatrix}0'_{\tilde{N}-N} & \mathbb{1}'_N\end{bmatrix} A\boldsymbol{z} = m$ as $v' = \begin{bmatrix}0'_{\tilde{N}-N} & \mathbb{1}'_N\end{bmatrix}$ constitute the kernel of M'. If we restrict the norm to the part $\Delta \log \boldsymbol{w}$ of the vector $\Delta \boldsymbol{z}$, Corollary 5.3.1 gives, considering $U = \begin{bmatrix}0_{N\times(\tilde{N}-N)} & I_{N\times N}\end{bmatrix}$, the estimator satisfying the additional constraint $S'\boldsymbol{z} = m$, such that $U\,\mathrm{Ker}(M')$ and $\mathrm{Ker}(S'_r)$ (with $S' = S'_r U$) are orthogonal with respect to the norm $\|\cdot\|_{D^2_w}$. We thus obtain that to minimize the relative error with respect to $\boldsymbol{w}$, one must find the BLUE estimator of $\boldsymbol{z}$ with the additional constraint:

$$\begin{bmatrix}0'_{\tilde{N}-N} & \mathbb{1}'_N\end{bmatrix} U' D^2_w U \boldsymbol{z} = \mathbb{1}'_N D^2_w \log \boldsymbol{w} = \mathbb{1}'_N D^2_{w^p} \log \boldsymbol{w}^p$$

The theoretical estimator satisfying this constraint among the solutions of the weighted least squares problem, i.e., $MV^\dagger M'\hat{\boldsymbol{z}} = MV^\dagger \log \boldsymbol{R}$, would be the one minimizing the approximation of $\frac{\|\Delta \boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|}$. This means that the linear unbiased estimator $\hat{\boldsymbol{z}}$ approximately minimizing $\frac{\|\Delta \boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|}$ lies on the line $MV^\dagger M'\hat{\boldsymbol{z}} = MV^\dagger \log \boldsymbol{R}$.

6.3.4 Minimum $\sin(\hat{\boldsymbol{w}}, \boldsymbol{w})$ error estimator

The last error measure introduced is the angle between the vectors $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$ and more precisely the absolute value of the angle sine. For a given estimator, this error will be the same for $\boldsymbol{w}^p$ or any $\boldsymbol{w} = \alpha \boldsymbol{w}^p$ since this measure is invariant to the scale of both vectors as shown in (6.1). This expression can be reduced to the previous results thanks to the trigonometrical formula of the sine (and as mentioned in [HOS20]):

$$|\sin(\hat{\boldsymbol{w}}, \boldsymbol{w})| = |\sin(\hat{\boldsymbol{w}}, \boldsymbol{w}^p)| = \inf_{\alpha\in\mathbb{R}} \frac{\|\alpha\hat{\boldsymbol{w}} - \boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|}$$

The problem of minimizing this error reduces thus to minimize $\frac{\|\Delta \boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|}$ for any particular $\boldsymbol{w}^p$ and the minimum (among all linear unbiased estimators $\hat{\boldsymbol{z}}$) of the approximation of $\frac{\|\Delta \boldsymbol{w}^p\|}{\|\boldsymbol{w}^p\|}$ lies on the line $MV^\dagger M'\hat{\boldsymbol{z}} = MV^\dagger \log \boldsymbol{R}$.

Part IV

Analysis, Results and Experimentation

The problem is to recover item weights based on models extending the BTL one. Those models are such that the weights $\boldsymbol{w}$ are defined up to a constant multiplicative factor. In other words, $p(\boldsymbol{w}) = p(\alpha\boldsymbol{w})$ where $p(\boldsymbol{w})$ represents the probability a judge ranks the items having scores $\boldsymbol{w}$ in a certain preference order. This means that if we consider a particular model, knowing the rankings probabilities, we know that the logarithm of the true weights define a line $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$ where $z = \left[(\log \boldsymbol{\theta})' \quad (\log \boldsymbol{w})'\right]'$.

A unique vector $\boldsymbol{z}$ on that line can be found by adding a constraint of the form $S'\boldsymbol{z} = m$. We have shown in previous Part that the vector has a best linear unbiased estimator $\hat{\boldsymbol{z}}$ in that constrained space $C = \{x | S'x = m\}$ (or equivalently a linear unbiased estimator minimizing a particular weighted $2-$norm $\|\boldsymbol{z} - \hat{\boldsymbol{z}}\|_{2,A}$). Those estimators lie on a line $MV^{\dagger}M'\hat{\boldsymbol{z}} = MV^{\dagger}\log \boldsymbol{R}$ and minimize $\log \boldsymbol{R} - M'\boldsymbol{z}$ in the covariance (pseudo-inverse) weighted least squares sense.

This Part aims to analyze the performances of our method regarding the existing ones. It is furthermore interesting to study the relative performances of the weighted least squares method regarding the different models used.

To do so, we will first analyze how close the estimators belonging to the set $MV^{\dagger}M'\hat{\boldsymbol{z}} = MV^{\dagger}\log \boldsymbol{R}$ are to the set of the true weights defined by $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$. This will be done in Chapter 7 by using the error measure $\sin(\boldsymbol{w}, \hat{\boldsymbol{w}})$. This error is indeed independent of the points $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ considered since the angle between $\alpha\boldsymbol{w}$ and

$\hat{\alpha}\hat{\boldsymbol{w}}$ does not depend on α and $\hat{\alpha}$:

$$\cos(\alpha\boldsymbol{w}, \hat{\alpha}\hat{\boldsymbol{w}}) = \frac{\langle \alpha\boldsymbol{w}, \hat{\alpha}\hat{\boldsymbol{w}} \rangle}{\|\alpha\boldsymbol{w}\|\|\hat{\alpha}\hat{\boldsymbol{w}}\|} = \frac{\alpha\hat{\alpha}\langle \boldsymbol{w}, \hat{\boldsymbol{w}} \rangle}{\alpha\hat{\alpha}\|\boldsymbol{w}\|\|\hat{\boldsymbol{w}}\|} = \frac{\langle \boldsymbol{w}, \hat{\boldsymbol{w}} \rangle}{\|\boldsymbol{w}\|\|\hat{\boldsymbol{w}}\|} = \cos(\boldsymbol{w}, \hat{\boldsymbol{w}})$$

This Chapter also studies the way one must perform an experiment in order to have an estimation of the scores as close as possible to the theoretical true weights of the items. This is done on several analyses based on the graph structure, the number of comparisons between items and the number of items to consider. Eventually, a comparison of the different methods allowing to conclude on the amount of information needed for having a given error regarding the model followed by the observation is performed.

Chapter 8 contains a movie application of our method. Based on user's scores, ranking of movies can be obtained, assuming that a user who prefers a movie will give it a higher score in opposition to movies he less liked. Assuming the populations preferences follow a particular model, our method can then be applied to recover global movie scores. The Chapter aims to point the differences in the associated global rankings of the studied movies out regarding the model used (considering pairwise comparisons with/without ties or considering triple comparisons between movies).

Chapter 7
Weighted Least Squares Method general performances

In the previous Part we saw that theoretically, the non-biased linear estimators minimizing the errors with the true weights of a probabilistic model lie on the line $MV^\dagger M\hat{\boldsymbol{z}} = MV^\dagger \log \boldsymbol{R}$, defining estimators of $\boldsymbol{w}$ up to a multiplicative factor. This Chapter aims to evaluate the accuracy of those estimators regarding the true randomly artificially generated value of the weights, regardless of the multiplicative factor, and this can be done using the sine error measure. Indeed, the angle between a vector $\boldsymbol{w}$ (from a vector $\boldsymbol{z}$ of the line $\log \boldsymbol{R} = M'\boldsymbol{z} + \boldsymbol{\epsilon}$) and a vector $\hat{\boldsymbol{w}}$ (from a vector $\hat{\boldsymbol{z}}$ of the line $MV^\dagger M\hat{\boldsymbol{z}} = MV^\dagger \log \boldsymbol{R}$) does not change if the vectors are multiplied by a scalar as shown in (6.1).

Remember that the sine between those two vectors is defined by

$$|\sin(\hat{\boldsymbol{w}}, \boldsymbol{w})| = \inf_{\alpha \in \mathbb{R}} \frac{\|\alpha\hat{\boldsymbol{w}} - \boldsymbol{w}\|}{\|\boldsymbol{w}\|}$$

And from Chapter 6, we found that

$$\begin{aligned}\min_{\hat{\boldsymbol{w}}} \mathbb{E}\left[\frac{\|\Delta \boldsymbol{w}\|}{\|\boldsymbol{w}\|}\right] &\approx \min_{\hat{\boldsymbol{w}}} \mathbb{E}\left[\|\Delta \log \boldsymbol{w}\|_{D_w^2}\right] = \min_{\hat{\boldsymbol{z}}} \mathbb{E}\left[\|U\Delta \boldsymbol{z}\|_{D_w^2}\right] \\ &= \min_{\hat{\boldsymbol{z}}} \sqrt{\operatorname{Tr}\left(D_w U \mathbb{V}\left[\Delta \boldsymbol{z}\right] U' D_w\right)}\end{aligned}$$

where $D_w = \operatorname{diag}\left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right)$ and $U = \begin{bmatrix}0_{N\times(\tilde{N}-N)} & I_{N\times N}\end{bmatrix}$. Assuming that the observations $\log \boldsymbol{R}$ have a well-defined variance, the minimum value corresponds to the one obtain with the estimator $\hat{\boldsymbol{z}}$ found using our weighted least squares method with an additional constraint $S' = \begin{bmatrix}0 & \mathbb{1}_N'\end{bmatrix} U' D_{\boldsymbol{w}}^2 U$.

All of this is based on a Taylor approximation. Figure 7.1 shows, for the Rao-Kupper model, that this estimation follows closely the true measure of the error computed with the sine as explained before. The results are similar for the other models, as shown in Appendix E on Figure E.2.

Since $\mathbb{V}[\Delta z]$ varies with $1/k$, we expect results evolving with $1/\sqrt{k}$ for the sine error measure (or the norm of the error on $\boldsymbol{w}$ or $\boldsymbol{z}$). Our graphs will therefore be shown in log-scale in order to be able to observe if $\log|\sin(\hat{\boldsymbol{w}}, \boldsymbol{w})| \approx -\frac{1}{2}\log k + c$ with an arbitrary constant c.

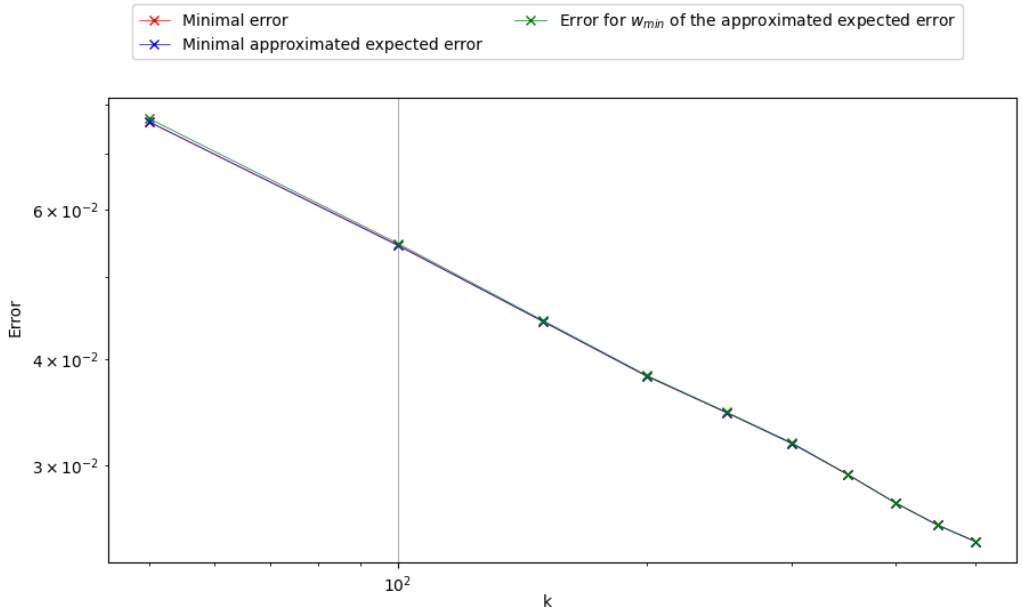


Figure 7.1: Illustration of error measures of the item weights for the Rao-Kupper model ($\theta = 1.5$). The red curve shows the sine error (minimal error obtained between the true weights $\boldsymbol{w}$ and its estimation $\hat{\boldsymbol{w}}_{min}$ belonging to the set $MV^{\dagger}M\hat{\boldsymbol{z}} = MV^{\dagger}\log \boldsymbol{R}$); the blue curve shows the error between the vectors $\log \boldsymbol{w}$ and $\log \hat{\boldsymbol{w}}$ for an estimator minimizing the weighted 2−norm with weights D_w; the green curve shows the error between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$. The experiments have been done on a complete graph of 10 items with a constant number of comparisons k between comparison sets; scores are generated on 100 simulation such that $\boldsymbol{w} \in [1, 20]$.

7.1 General methods comparison

The first analyses will be done on complete graphs of 10 items, meaning that an object is compared to all the others. We furthermore consider that the number of observations is the same for all the comparisons, namely that for an event $*$ it holds that $k_* = k$. Scores are generated randomly between 1 and 20. Each result is then obtained by computing an average on 100 scores simulations.

7.1.1 Weighted Least Squares methods comparison

The weighted least squares method requires the knowledge of the weight matrix. As mentioned in Chapter 4, this weight matrix can be chosen as the pseudo-inverse of the covariance matrix of $\log \boldsymbol{R}$, or the pseudo-inverse of the diagonal matrix of the variances of $\log \boldsymbol{R}$. Those matrices are unknown in practice, but we saw different manners to estimate them.

We can first approximate those matrices using a Taylor development, in which case we obtain the matrices V and D_V. To compute the matrices, the true weights $\boldsymbol{w}$ of the items are needed, as they depend on the different probabilities to prefer a particular ranking. Those values are unknown in practice, but we can approximate them with the corresponding observed values, i.e., the fraction of times the ranking was observed, in

which case we obtain the matrices $\hat{V}$ and $\hat{D}_V$. It indeed holds that for an event $*$

$$p_* \underset{k_* \to \infty}{\to} h_*$$

On Figure 7.2, the error obtained on the estimator $\hat{\boldsymbol{w}}$ using our weighted least squares method on the Rao-Kupper model with V, $\hat{V}$, D_V and $\hat{D}_V$ is shown where the matrices are computed using the true and the empirical Taylor approximation of the covariance matrix and the diagonal matrix of variance of $\log \boldsymbol{R}$.

Another possibility to approximate the two matrices is to use an iterative method as presented in Chapter 4. This is also shown on Figure 7.2 for the Rao-Kupper model.

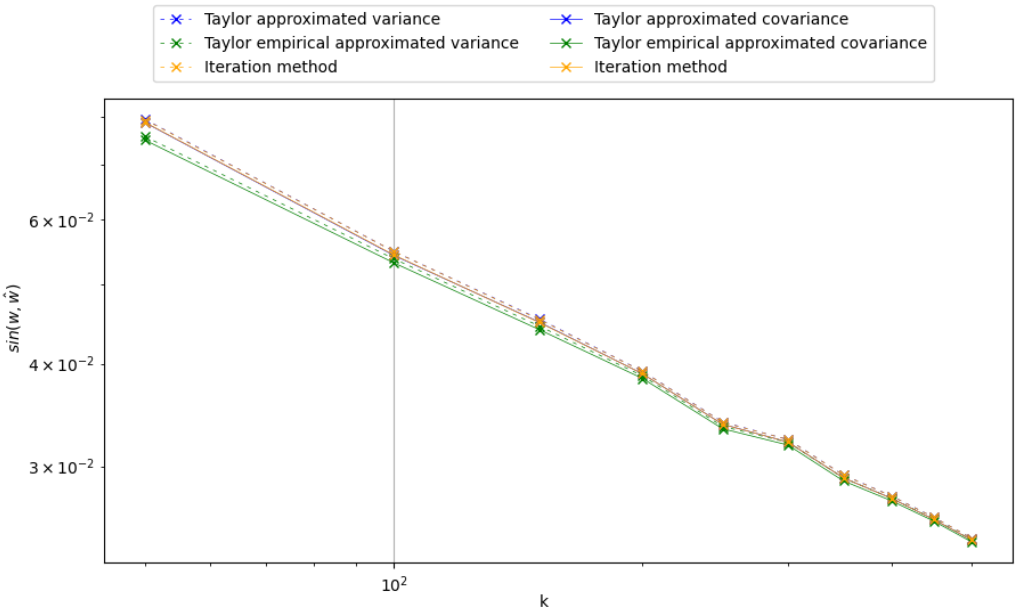


Figure 7.2: Comparison of the performances obtained for the weighted least squares method with different weight matrices for the Rao-Kupper model ($\theta = 1.5$).

We observe surprisingly that the weighted least squares method with $\hat{V}$ (covariance matrix computed using the approximated Taylor expansion) performs better than the method using V (covariance matrix computed with the true Taylor expansion of the expressions). This was also observed for the BTL model in [HOS20].

We have shown theoretically in Chapter 6 that the covariance matrix leads in expectation to better results than the diagonal matrix of the variances. Globally, in practice, the results using both approaches for the weight matrix are similar. Performances for the methods using the covariance matrix stay however slightly better as shown on the Figure, which confirms the theoretical results. This stays true for all the models except for the Plackett-Luce model, where a weighted least squares method using $\hat{D}_V$ (diagonal matrix of the variance computed with the Taylor empirical approximation of $\log \boldsymbol{R}$) has better performances compared with the method using $\hat{V}$ computed the same way as shown in Appendix E on Figure E.3.

7.1.2 Methods comparison

As we have seen in the previous Section, the weighted least squares method used to recover item weights performs globally better when using the covariance matrix of $\log \boldsymbol{R}$ as weight matrix. This will thus be used to compare our method with the existing ones.

The literature proposes actually few other methods allowing the recovering of item weights based on ranking preference observations. Several methods were introduced in Chapter 3 and will be compared to our method in this Section.

A widespread method is the **maximum likelihood** one, which consists of finding the vector $\boldsymbol{w}$ for which the probability of observing the global ranking associated to that vector is the highest. This method is an iterative method and has been developed for all the models studied in Chapter 3. This means that an initial vector $\begin{bmatrix}\boldsymbol{\theta}^0 & \boldsymbol{w}^0\end{bmatrix}$ must be determined. The method is sensitive to that candidate, and we opted for the following initial vectors for the different methods:

$$w^0_{\mathrm{BTL},i} = K^1_i \qquad w^0_{\mathrm{MBT},i} = 2K^1_i + K^2_i \qquad w^0_{\mathrm{PL},i} = K^1_i + K^2_i$$
$$w^0_{\mathrm{RK},i} = K^1_i + K^=_i \qquad \theta^0_{\mathrm{RK}} = 2K^= \qquad w^0_{\mathrm{D},i} = K^1_i + \frac{K^=_i}{2} \qquad \theta^0_{\mathrm{D}} = K^=$$

where K^j_i denote the total number of times the object i has rank j in the observations and $K^=$ is the total number of ties among the observations. Those initial points have been chosen regarding the methods in order to have a starting weight vector depending on the rank of each item regarding the observations.

The second presented method was proposed in [GSK69]. The method presented is a **weighted least squares method** for the Bradley-Terry-Luce, Davidson, Rao-Kupper and Mallows-Bradley-Terry models. Their method amounts to our method in the specific case of a constraint $\log w_N = 0$ when the considered equations are the same, i.e., the same ratio $R = h_\alpha/h_\beta$. Note that the way the method is presented does not allow us to generalize it to other models, or at least less easy than what our method can offer. Besides, they only consider the same equations as our method for the BTL and Rao-Kupper models.

The last methods consist first of the **least squares** method without additional weights and on the other hand of a method transforming the observations into pairwise comparisons observations without ties. The first method is self-explanatory, and the latter method then applies a **weighted least squares method on a BTL transformed model** [HOS20] as described in Chapter 3.4. Remember that for pairwise comparisons models with ties, 2 different ways have been presented in Chapter 4 to transform the observations, namely ignoring the ties or splitting the ties observations between the two items (half of them goes to each item). Figure 7.3 shows the performances for the two pairwise comparison models, allowing ties on the weighted least squares method. We observe that no method stands out since, for Davidson, the method which ignores ties performs better while the other one provides better results for the Rao-Kupper model.

Those results are indeed what is expected. For the method which consist of ignoring

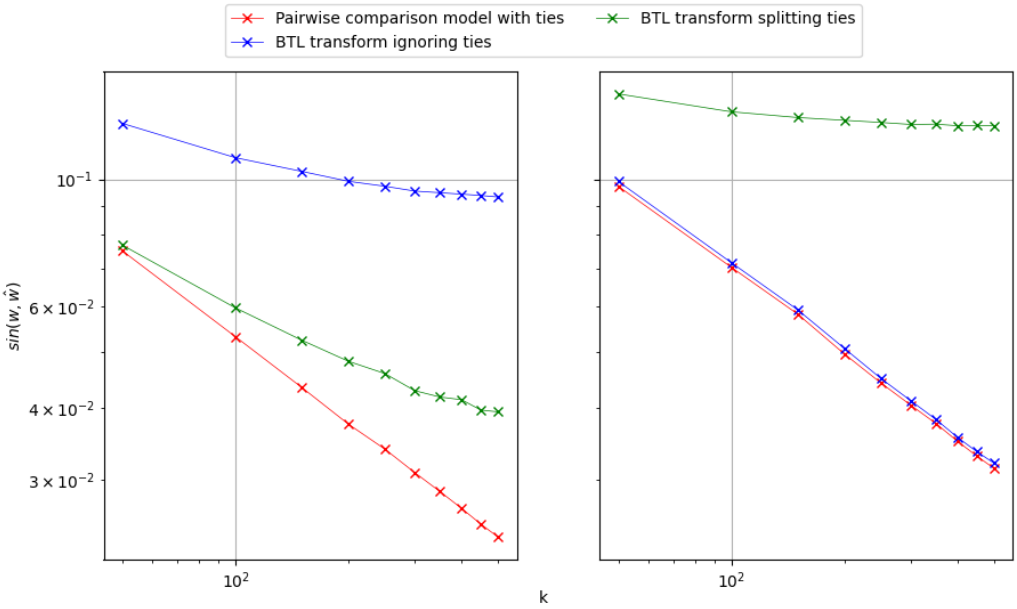


Figure 7.3: BTL transform method for the pairwise comparisons models with ties, Rao-Kupper ($\theta = 2$) on left and Davidson ($\theta = 1$) on right.

tie observations we obtain the new ranking probabilities for the two models

$$(\tilde{p}_{i,j})^{\mathrm{RK}} = \frac{p_{i,j}^{\mathrm{RK}}}{p_{i,j}^{\mathrm{RK}} + p_{j,i}^{\mathrm{RK}}} = \frac{w_i(w_j + \theta w_i)}{w_i(w_j + \theta w_i) + w_j(w_i + \theta w_j)}$$
$$(\tilde{p}_{i,j})^{\mathrm{D}} = \frac{p_{i,j}^{\mathrm{D}}}{p_{i,j}^{\mathrm{D}} + p_{j,i}^{\mathrm{D}}} = \frac{w_i}{w_i + w_j}$$

The ratio $R_{i,j} = h_{i,j}/h_{j,i} \approx p_{i,j}/p_{j,i}$ which is approximately equal to w_i/w_j for the BTL model becomes then

$$R_{i,j}^{\mathrm{RK}} \approx \frac{(\tilde{p}_{i,j})^{\mathrm{RK}}}{(\tilde{p}_{j,i})^{\mathrm{RK}}} = \frac{w_i(w_j + \theta w_i)}{w_j(w_i + \theta w_j)}$$
$$R_{i,j}^{\mathrm{D}} \approx \frac{(\tilde{p}_{i,j})^{\mathrm{D}}}{(\tilde{p}_{j,i})^{\mathrm{D}}} = \frac{w_i}{w_j}$$

which explains the good results for the Davidson model BTL transform using this method.

The second proposed method consists of sharing the tie observations between the two items of the set. In this case the new ranking probabilities become

$$(\tilde{p}_{i,j})^{\mathrm{RK}} = p_{i,j}^{\mathrm{RK}} + \frac{p_{i=j}^{\mathrm{RK}}}{2} = \frac{w_i(w_j + \theta w_i)}{(w_j + \theta w_i)(w_i + \theta w_j)}$$
$$(\tilde{p}_{i,j})^{\mathrm{D}} = p_{i,j}^{\mathrm{D}} + \frac{p_{i=j}^{\mathrm{D}}}{2} = \frac{w_i + \theta\sqrt{w_i w_j}/2}{w_i + w_j + \theta\sqrt{w_i w_j}}$$

The ratio $R_{i,j} = h_{i,j}/h_{j,i} \approx p_{i,j}/p_{j,i}$ are then approximated by

$$R_{i,j}^{\mathrm{RK}} \approx \frac{(\tilde{p}_{i,j})^{\mathrm{RK}}}{(\tilde{p}_{j,i})^{\mathrm{RK}}} = \frac{w_i}{w_j}\frac{(1+\theta^2)w_j + 2\theta w_i}{(1+\theta^2)w_i + 2\theta w_j}$$
$$R_{i,j}^{\mathrm{D}} \approx \frac{(\tilde{p}_{i,j})^{\mathrm{D}}}{(\tilde{p}_{j,i})^{\mathrm{D}}} = \frac{w_i + \theta\sqrt{w_i w_j}/2}{w_j + \theta\sqrt{w_i w_j}/2}$$

We observe that for $\theta \approx 1$, it holds that $R_{i,j}^{\mathrm{RK}} \approx w_i/w_j$ as for the BTL model since $1 + \theta^2 \approx 2\theta$. This explains that this method gives better results for the Rao-Kupper model. Note that the higher the parameter θ is, the higher the gap between the performances of the Rao-Kupper model and this BTL transform method on the model for a weighted least squares method will be.

Figure 7.4 shows the sine error evolution in function of the number of observations per subset of items for the above methods for the Rao-Kupper model. The results are similar for the other models, as shown in Appendix E on Figure E.5. For the BTL transform method, we only kept the one with the best performances for the Rao-Kupper and Davidson models.

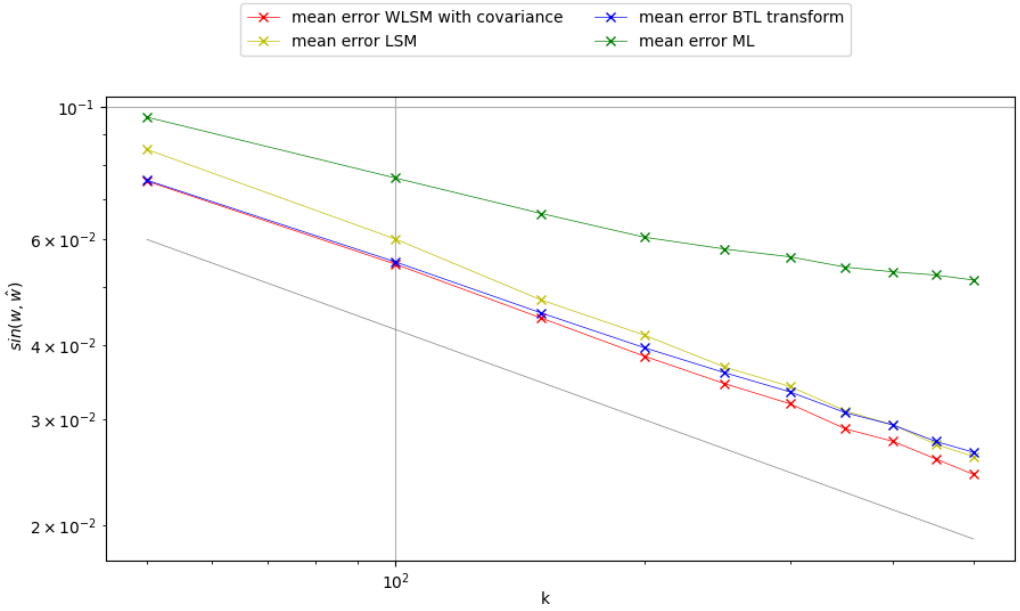


Figure 7.4: Comparison of the different methods for the Rao-Kupper model ($\theta = 2$). The gray curve represents the line $\log|sin(\hat{\boldsymbol{w}}, \boldsymbol{w}| = -1/2 \log k + c$ for an arbitrary constant c.

First, observe that the maximum likelihood method performs not as good as the other ones. This might be because the initial guess is not optimal. The other methods do not depend on an initial vector choice, which is a major asset. We also observe that the BTL transform method performs quite good compared to the others.

Furthermore, the weighted least squares methods performs better than the least squares method. The gap between the errors of the two methods increases when the weights values vary more, namely when the ratio between the highest and the smallest score increases. An illustration for weights $\boldsymbol{w} \in [1, 5]$ and $\boldsymbol{w} \in [1, 1000]$ is given in

Appendix E on Figure E.4 to observe the impact of item scores on the difference between the least squares and the weighted least squares methods.

7.2 Models comparison

It is also interesting to analyze the difference in the performances of the different models for our weighted least squares method. In other words, we will analyze for the same amount of information, which model will perform better.

The analyses will be done on 10 items, considering all the possible sets of items (pairs or triplets). Scores are generated randomly between 1 and 20. Each result is then obtained by computing an average on 100 scores simulations.

We assume that the amount of information given to a model corresponds to the memory allocation of the observations. More specifically, it holds that the information corresponding to an observation of a triple comparison needs a memory allocation of $\log_2 6$ bits while for pairwise comparisons with BTL, $\log_2 2 = 1$ bit is sufficient and $\log_2 3$ bits are needed for pairwise comparisons observation with ties.

Figure 7.5 illustrates the performances for the 5 models studied in this report. The values of k are the number of observations done between two items for the BTL model. The corresponding number of observations for pairwise comparison models with ties and for triple comparison models are given by[1]

$$k_{\mathrm{RK,D}} = \frac{k_{\mathrm{BTL}}}{\log_2(3)} \qquad k_{\mathrm{PL,MBT}} = \frac{3k_{\mathrm{BTL}}}{(N-2)\log_2(6)}$$

We observe that for a same amount of information, the BTL model will perform better than the other models. We also observe that the Plackett-Luce model has an error which decreases faster than the error for the other models. If one may choose the type of observation collected, opting for pairwise comparison without ties would allow in theory to reach the best performances for a same amount of information.

7.3 Graph structure and number of comparisons

In the above analyses, we supposed that we had observations for a complete graph (all objects compared to each other) and that the number of observations on a set of n items was constant. This is rarely the case in practice, and this Section aims to observe the impact of the graph structure and the variation of the number of observations per item set on the performances of our weighted least squares method.

For all the experiments of this Section, the scores are generated randomly between 1 and 20. Each result is then obtained by computing an average on 100 scores simulations.

[1]The total amount of information is given by $K = kn_s$ where k is the number of ranking observations available per set of n items and $n_s = \frac{N!}{n!(N-n)!}$ is the number of different sets of n items one may observe on a complete graph of N items

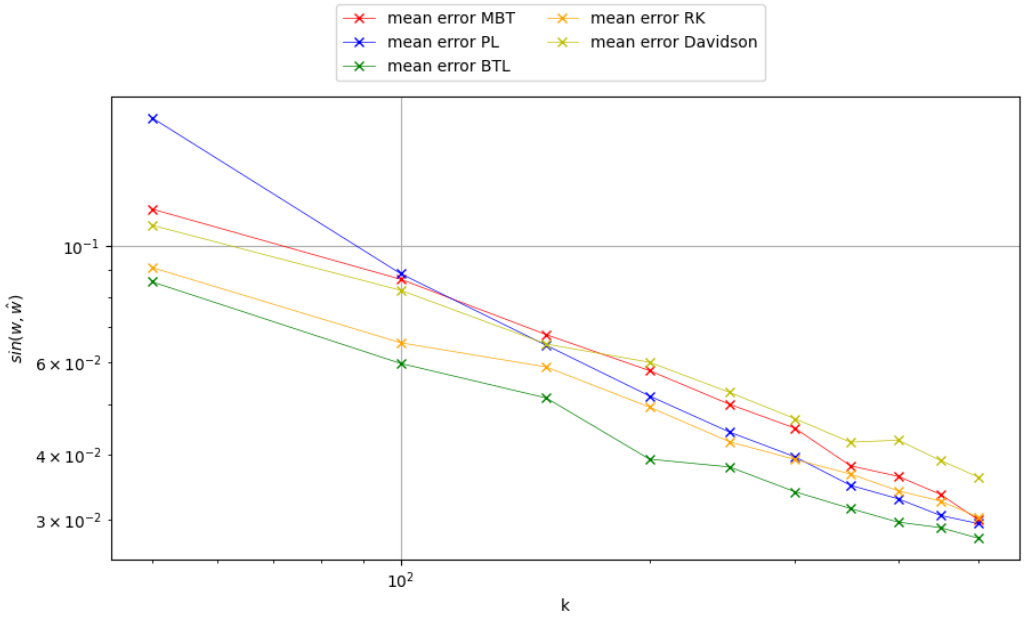


Figure 7.5: Performances obtained for the recovering of weights using our weighted least squares method on the different models ($\theta^{\mathrm{RK}} = 1.5, \theta^{\mathrm{D}} = 0.5$).

7.3.1 Allocation of the number of comparisons

From a circle to a complete graph We are interested in graphs with a variable number of comparisons between the items sets. We will analyze the performances of our method on a complete graph of 10 items with small number of comparisons between non-neighbors items. The analysis will be performed on a graph with a circle structure, and the structure will then evolve to increase progressively the number of comparisons between non-neighbors nodes until we end up with a complete graph with a constant total number of comparisons. This is illustrated on Figure 7.6.

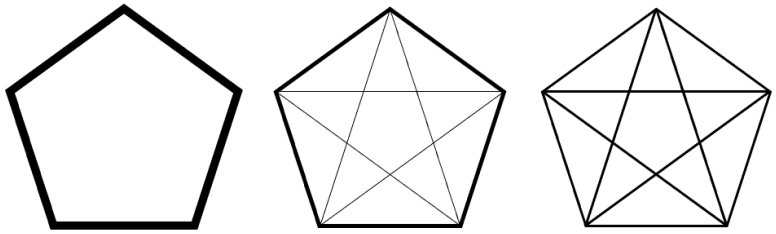


Figure 7.6: Illustration of the concept described: from a circle graph structure, the number of comparisons decreases progressively on the border edges and increases progressively between non-neighbors nodes until all the edges contains the same number of comparisons.

Note that all the graph structures considered have an equivalent amount of information regarding all the observations. The result for the above experiment is given on Figure 7.7 for the Rao-Kupper model. The underneath analysis is similar for the different models as shown in Appendix E on Figure E.6.

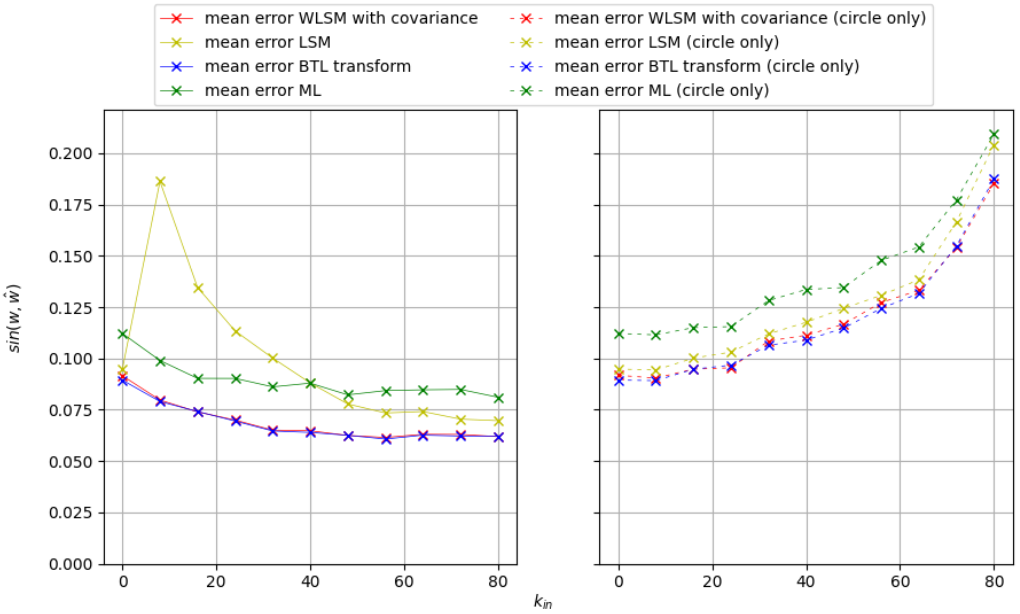


Figure 7.7: Performances of the different methods applied on the Rao-Kupper model ($\theta = 2$) for the experiment described above. The parameter k_{in} denotes the number of comparisons made between non-neighbors' nodes on the inside of the graph structure. The progressive increase of k_{in} hides the decrease of k_{out} between neighbor nodes (on the outside circle graph). The left figure shows the performances on the complete graph, while the right graph contains the results obtained on the circle graph structure only.

We observe that our least squares method gives the best performances. The error obtained when considering only the observations between neighbor nodes increases with the number of information moved to centered edges and for the least squares method applied on the complete graph, the error first increases before decreasing progressively. Those results are indeed what is expected. The increasing of the error for the least squares method stems on the fact that the information contained in the center edges is less accurate compare to comparisons between neighbor items. This is due to the small number of comparisons between non-neighbors nodes, in other words, the fraction of times a ranking is observed does not approximate the probability of observing that ranking accurately due to the small value of k. The results obtained on the circle graph are also expected, since the amount of information available on the circle decreases due to the progressive decrease of k.

This experiment is interesting since it informs us that having a complete graph with a low number of information is better than having a non-complete graph structure with a lot of observation between the compared items. This indeed holds since the left curves of Figure 7.7 are decreasing. It also holds that one must keep all the available information even if the number of comparisons between items is low compared to other comparisons. Except for the least squares method, all the item weight recovering methods take indeed the amount of information per comparisons into account.

Increasing comparisons or item number If we are interested in a certain number of items, one could ask if it is better to increase the number of comparisons between the actual items or to consider extra items to increase the accuracy of the estimated

scores of our items of interest, as illustrated on Figure 7.8.

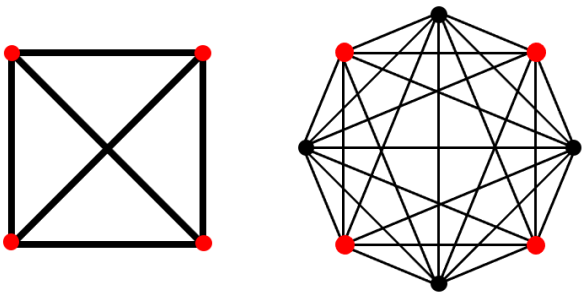


Figure 7.8: Illustration of the concept described: if one is interested in the score of 4 items is it better to have more information on the graph with only those items or to have less information on that graph but a higher number of items.

The analysis will be performed on a set of 5 items of interest. The first experiment keeps the number of items constant and progressively increases the number of comparisons (equal for all the performed comparisons), starting from $k = 50$. The second one keeps the number of comparisons per subset of compared items constant with $k = 50$ and progressively increases the number of items from $N = 5$ up to $N = 10$.

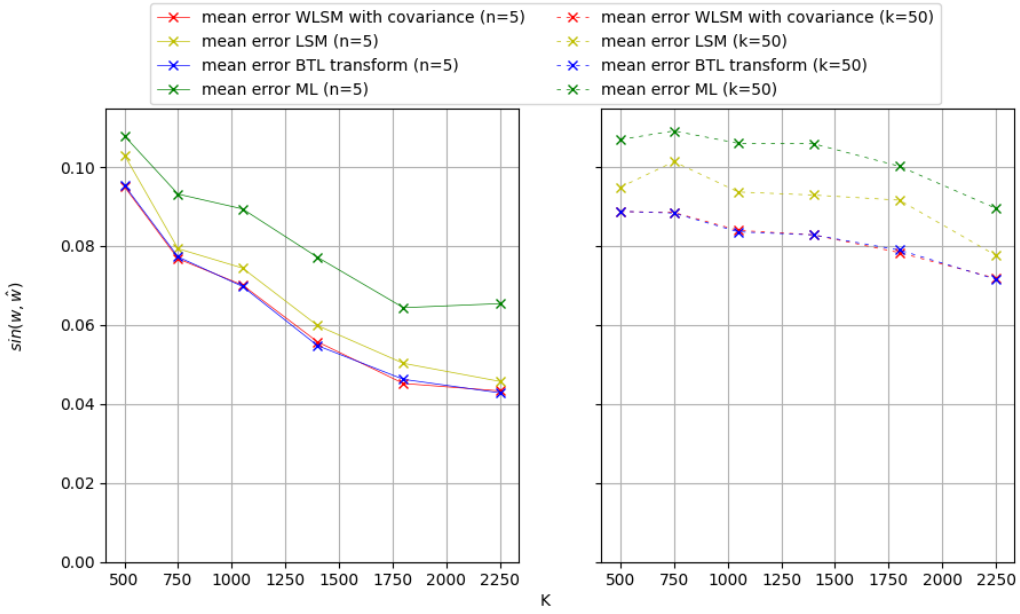


Figure 7.9: Performances of the different methods applied on the Rao-Kupper model ($\theta = 2$) for the experiment described above. The parameter K denotes the amount of information contained in the graph, namely the number of ranking observations per pair of items multiplied by the number of pairs in the graph. The left figure shows the performance on the graph of 5 nodes with an increasing number of observations between the items while the right graph shows the results for a constant number of observations between item ($k = 50$) for an increasing number of items.

Figure 7.9 shows the accuracy of those two approaches for a same total amount of information on the Rao-Kupper model. The performances for the other models are

given in Appendix E on Figure E.7. Let K be the amount of available information of the experiment. On a graph of N nodes for ranking observations of n items, it holds that

$$K = kn_s = k\frac{N!}{n!(N-n)!}$$

where n_s is the number of sets of n items in the graph structure considered (the last equality only holds in the case of a complete graph structure). The experiment is done such that K is the same on a graph on $N = 5$ nodes and on a graph with a number of observations between items of $k = 50$. We observe that it is more advantageous to put the information between the nodes of interest compared to adding comparison with non-interesting nodes. The performances on the graph of 5 nodes are indeed better than the one for a constant number of observations between more items.

7.3.2 Graph structure

We decided to analyze the impact of the graph structure by looking at the performances on 4 different structures composed of 10 items considering the same number of comparisons per subset of compared items. The results are shown on Figure 7.10 for the Rao-Kupper model with the following graph structures: line, circle, 2D grid of size 2×5 and Erdös-Rényi with a probability $\log(N)/N$ to connect 2 nodes[2]. The other models show similar results, as it can be seen on Figures E.8, E.9 and E.10 in Appendix E (for the triple comparison models, the graph structure have been adapted as shown on Figure E.1).

We observe that the results are similar for all the structures, even though the performances are better when there are more connections between the items for a same number of comparisons by subsets of items (pairs/triplets), probably because the total amount of information is higher. It is indeed noticeable that when the number of connections between the items is low, the BTL transform method performs better than our weighted least squares method.

[2]The Erdös-Rényi graph is a randomly generated graph where each edge belongs to the graph with a probability p, independently of the other edges.

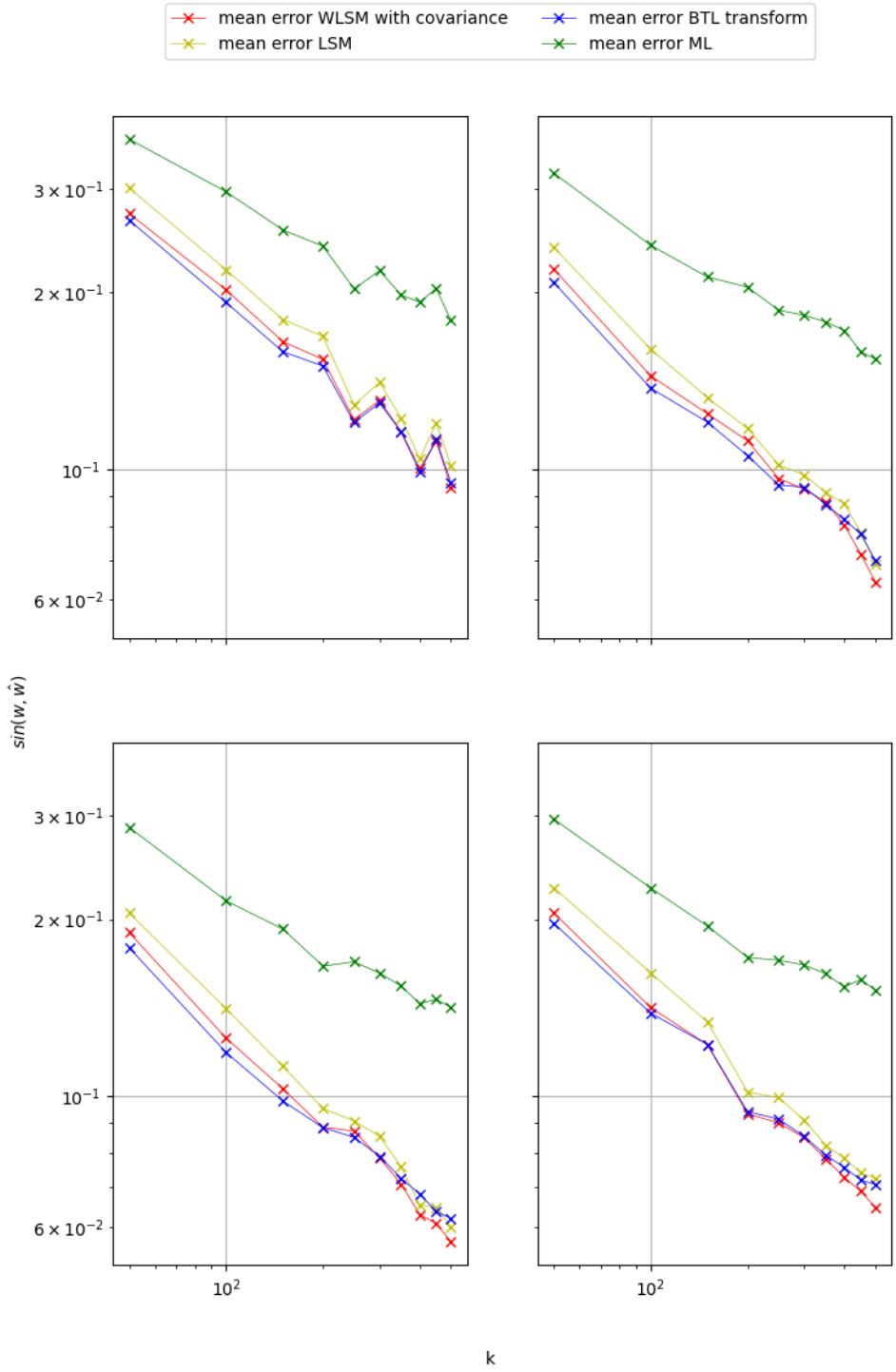


Figure 7.10: Performances of the Rao-Kupper model ($\theta = 2$) for different graph structures. The upper left graph shows the results on the line, the results on a circle are on the upper right figure, the lower left graph shows the performances on a 2D grid structure and the results for an Erdös-Rényi graph structure the results are shown on the lower right figure.

Chapter 8
Experimental application: Netflix movies ranking

In this Chapter, our aim is to show how our method can be used for a practical application, namely ranking some movies. The dataset used comes from the Netflix Prize competition[1] opened from 2006 to 2009. Those data consist of integer user ratings between 1 and 5 of several movies.

The purpose is to find different ways of getting the global merit of each movie and a global movies ranking. Based on user's ratings, a straightforward method is to compute the average rating of each movie and to sort the movies by decreasing order of their average rating. The fact is that, with user ratings, the rating scale is specific to each user. In other words, one can give ratings only between 3 and 5 while another one could never give a rating of 5. All of this depends on the user's perception of the merit values.

Movies preferences are more comparable between users. A rating of 3 from one user and a rating of 3 from another user does not always represent the same merit since it is an absolute value, but a user preference between movie A and B is relative and without a numerical value.

We chose 15 famous movies given in Table 8.1 with a corresponding number given by their ranking in the average score ranking. The rating data from approximately 50 000 users from the Netflix dataset are transformed into preferences data by considering that a movie with a higher rate is preferred by the user. Based on those new data, we applied our weighted least squares method to find some model-specific scores and to recover a global ranking. Note that in the cases of BTL, MBT and PL models, ties observations are ignored. The results can be compared with those obtained with the average rating of each movie. As explained previously, those scores obtained with our method are defined up to a multiplicative factor, so we took the scores such that the mean score is equal to the mean of the average ratings.

Figure 8.1 shows the score distribution with the average ratings and the 5 comparison models scores. The average ratings vary between 3.5 and 4.7 while comparison models scores vary between 0.5 and 12 to 25 for the highest score, meaning that the ratio between the highest score and the lowest one is very different. We observe that the comparison models scores are different from each other, the score parameters are indeed intrinsic to each model.

[1] https://web.archive.org/web/20090924184639/http://www.netflixprize.com/, archive from the original website http://www.netflixprize.com/. Dataset found on https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

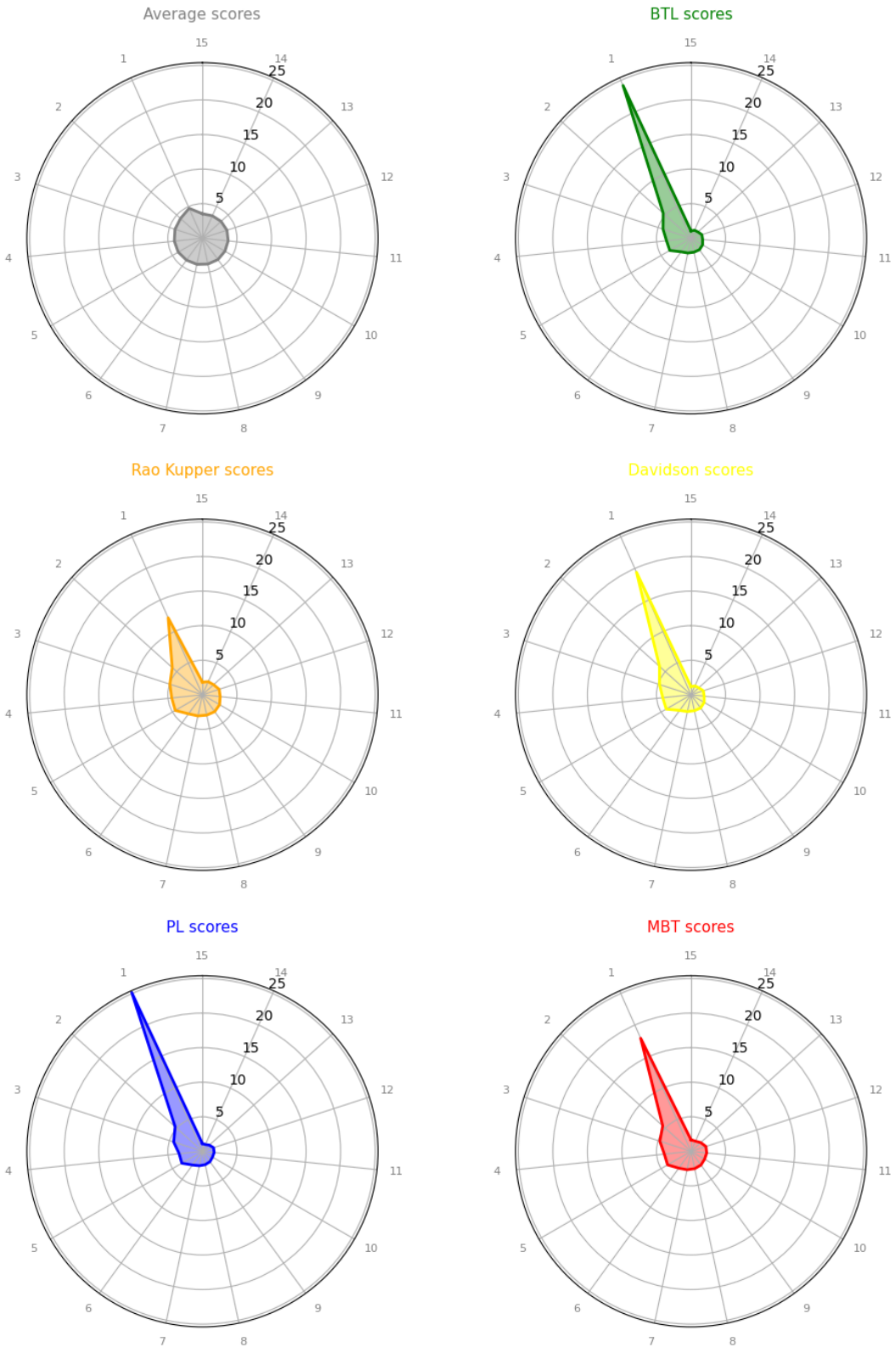


Figure 8.1: Representation of the scores attributed to the movies (from movie 1, "The Lord of the Rings", up to movie 15, "Mission: Impossible"). All the score distributions are constrained to have the same average score.

TOP	Movies	Average Rating
1	The Lord of the Rings	4.73
2	Pirates of the Caribbean	4.27
3	Back to the Future	4.17
4	Harry Potter	4.09
5	The Pianist	4.04
6	Jurassic Park	3.92
7	Top Gun	3.87
8	Love Actually	3.83
9	Grease	3.82
10	Titanic	3.76
11	Bridget Jones's Diary	3.74
12	Edward Scissorhands	3.73
13	Notting Hill	3.64
14	Star Wars	3.55
15	Mission: Impossible	3.54

Table 8.1: Chosen movies with their user's average rating. Each movie is attributed a number given by their rank in the ranking based on those given scores (from the highest to the lowest).

Based on those score results, we can generate a global TOP15 ranking of the movies and compare it to the average rating raking. Tables 8.2 and 8.3 shows the resulting ranking for the BTL / Rao & Kupper / Davidson models and Plackett-Luce / Mallows-Bradley-Terry models respectively.

Table 8.2 shows that for the BTL model, all movies have a different position in the ranking compared to the average ratings raking, but we can observe that the general trend is similar, the rank of each movie differs by 1 to 3 (except for one movie with 9). For ties models, both models give really similar results to each other: only the last pair of movies has exchanged positions. The rankings are similar to the BTL ranking. For triplet models, there are fewer differences with the average ratings raking: half of the movies have the same rank, and the others differ by one or two.

All of this shows indeed that scores parameters depend on the considered model, since each model has a different ranking. The general trend of the rankings is however similar, and this is probably due to the similar properties of each model, all of them extend the initial BTL model.

Note also that this application is based on movie rating of 1, 2, 3, 4 or 5. This means that a lot of ties in pairs, and even more in triplets, of movies occur and forcing us to ignore a part of the information for models not allowing ties. The two pairwise comparisons with ties models are thus based on the same amount of information than the average ratings ranking, and thus give the most relevant results.

Bradley-Terry-Luce		Rao & Kupper	Davidson	
Pirates of the Caribbean	▲ 1	The Lord of the Rings	● 0	
The Lord of the Rings	▼ 1	Pirates of the Caribbean	● 0	
Harry Potter	▲ 1	Harry Potter	▲ 1	
Back to the Future	▼ 1	Back to the Future	▼ 1	
Jurassic Park	▲ 1	Jurassic Park	▲ 1	
Top Gun	▲ 1	Top Gun	▲ 1	
Titanic	▲ 3	Titanic	▲ 3	
Grease	▲ 1	Grease	▲ 1	
Love Actually	▼ 1	Love Actually	▼ 1	
Bridget Jones's Diary	▲ 1	Bridget Jones's Diary	▲ 1	
Edward Scissorhands	▲ 1	Edward Scissorhands	▲ 1	
Notting Hill	▲ 1	Notting Hill	▲ 1	
Mission: Impossible	▲ 2	Mission: Impossible	▲ 2	
The Pianist	▼ 9	Star Wars ● 0	The Pianist	▼ 9
Star Wars	▼ 1	The Pianist ▼ 10	Star Wars	▼ 1

Table 8.2: Resulting rankings based on BTL, Rao & Kupper and Davidson model scores. The differences with the ranking based on the average rating are highlighted by ▲ for better ranked movies, ▼ for lower ranked movies and ● for identically ranked movies.

Mallows-Bradley-Terry		Plackett-Luce	
The Lord of the Rings			● 0
Pirates of the Caribbean			● 0
Back to the Future			● 0
The Pianist			▲ 1
Harry Potter			▼ 1
Jurassic Park			● 0
Top Gun			● 0
Love Actually			● 0
Grease			● 0
Bridget Jones's Diary			▲ 1
Titanic	▼ 1	Edward Scissorhands	▲ 1
Edward Scissorhands	● 0	Titanic	▼ 2
Notting Hill			● 0
Mission: Impossible			▲ 1
Star Wars			▼ 1

Table 8.3: Resulting rankings based on Mallows-Bradley-Terry and Plackett-Luce model scores. The differences with the ranking based on the average rating are highlighted by ▲ for better ranked movies, ▼ for lower ranked movies and ● for identically ranked movies.

Conclusion

The purpose of this master thesis was to propose a generalized weighted least squares method for BTL extended models extending the method developed in [HOS20]. Such a method aims at estimating item scores based on noisy observations of preference rankings. This method can be applied on several models, and this report focused on two pairwise comparisons models allowing ties in the observations and on two triple comparisons models without ties. Based on comparison observations, recovering item scores allows having a global ranking of the items based on their weights, and it would make it possible to predict future observations assuming that the observations indeed follow the model used to solve the problem. Furthermore, using observations following a given model, we compared the performances of the developed method regarding the existing ones.

Summary

We extended the weighted least squares method of the BTL model for a larger set of models. In other words, if the considered model is such that a function of the probabilities of observing the rankings is linear in the piecewise logarithm of the item weights, our weighted least squares method can be applied to recover the scores. The method uses a weight matrix equal to the pseudo-inverse of the observation covariance matrix, which in the case of the BTL model is also equal to the inverse of the variance matrix. Note that this matrix is unknown in practice, but it can be approximated in several ways. We noticed that using an empirical approximation of the covariance expressions provides low error on the score estimation.

Recovering item weights can indeed be seen as a problem of estimating parameters of a linear model based on noisy observations. Some results about Best Linear Unbiased Estimators already existed allowing us, with some conditions, to find the linear unbiased estimator minimizing the variance thus also the norm of the error. The hypotheses of the existing theorems were not satisfied by the linear system, and we managed to prove additional results for more general problems of the class. More precisely, our system had non-observable parameters, since the models were such that the scores defined the probabilities up to a scale factor. By constraining the space to a subspace containing only one possible vector of scores, it was then possible to find the BLUE estimator of the parameters in that subspace. Furthermore, the observations had a variance different from the identity matrix (correlated and/or non-homoscedastic observations). In this case, in order to find the BLUE of the true item weights, it was mandatory to reformulate the problem to end with a linear system respecting the previous hypothesis. This reformulation made the pseudo-inverse of the covariance of the observations appears.

By considering a unique vector of parameters, it was furthermore possible to deter-

mine the estimator minimizing a given 2−norm of the error. Note that the obtained results were proven optimal for $\log w$ and not for the vector of scores w. We managed however to link the two quantities, and we thus found theoretical results on the error of w. We also found theoretically the minimal value of this error.

We observed that among the methods described, our weighted least squares method empirically outperformed the others. This held for any graph structure studied. We also considered transforming the observations in pairwise comparisons exploitable by the BTL model. This method provided similar results for some models.

Eventually, we applied our method on a movie application. We aimed to provide a global ranking of a number of Netflix movies based on their estimated scores. The obtained results were different from the ones obtained by taking the mean user scores.

Research perspectives

Further research leads on this topic are various. Some ideas following this thesis are explained hereunder.

As mentioned above, our weighted least squares method uses the pseudo-inverse of the covariance matrix of the observations as the weight matrix. This matrix, unknown in practice, is approximated empirically in order to obtain results. We observed that the empirical approximated covariance matrix provides better performances than the matrix approximated with the true probabilities values (only available theoretically). A deeper study about the approximation of the covariance matrix and its impact on the performances of the error might allow us to better understand this observation. A first empirical analysis for the BTL model has already been studied in a previous master thesis [Win21].

Besides, regarding the experimental part of the research, a deeper analysis on the obtained performances considering the graph structure and the number of comparisons between the item is interesting to understand the impact of those parameters on the estimator error. This also includes a deeper understanding of the results obtained when one restricts its study to a small set of nodes from a larger set of items. We have indeed observed that having a lower number of comparisons between items provide better performances than having a structure with less connection between items but a higher number of observations between those items. After, we also deduced from our experiments that on a set of interesting items, it is preferable to provide more information about the connections between those items than to add items to the set providing additional information about the interesting items. In a more general point of view, one could then wonder how optimally allocate the possible comparisons?

As mentioned in the introduction, several methods exist when observations on global rankings are provided. The models studied in this report such as the Plackett-Luce or the Mallows-Bradley-Terry models may provide such global ranking probability models. It could be interesting to compare the performances of an extension of our weighted least squares method for observations on N items with those existing aggregated methods.

Future research could also aim to develop a method which updates item weights based on an prior vector of scores, for example using past observations. This could be applied for our experiment in order to update the weights if other judges went to provide additional observations.

A last lead we thought interesting is to use the item scores found by our weighted least squares method to predict user preferences based on the observations. In the case of global (not user-specific) preference probabilities, prediction outcome can already be computed based on the model probabilities and previous observations, such as for sport tournament, and this has already been studied in some research. User-specific preferences could also be predicted by those models. Our practical application was based on movie ratings by users, initially in the context of movie ratings prediction, and this can be extended to movie preferences prediction. By considering scores specific to an identified group of users, one might predict the user preference probabilities of this group. The matter would be the way of identifying groups of users.

Bibliography

[Bea77] Robert J Beaver. Weighted least-squares analysis of several univariate bradley-terry models. *Journal of the American Statistical Association*, 72(359):629–634, 1977.

[BG75] Robert J. Beaver and D. V. Gokhale. A model to incorporat within-pair order effects in paired comparisons. *Communications in Statistics*, 4:923–939, 1975.

[BLSR20] Heejong Bong, Wanshan Li, Shamindra Shrotriya, and Alessandro Rinaldo. Nonparametric estimation in the dynamic bradley-terry model. In *International Conference on Artificial Intelligence and Statistics*, pages 3317–3326. PMLR, 2020.

[BR73] Robert J Beaver and PV Rao. On ties in triple comparisons. *Trabajos de estadistica y de investigacion operativa*, 24(1):77–92, 1973.

[BT52a] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi:10.2307/2334029.

[BT52b] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: Ii. the method for blocks of size three. 1952.

[Cat12] Manuela Cattelan. Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science*, 27(3):412–433, 2012.

[Dav70] Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970.

[FMM22] Fiorenzo Franceschini, Domenico A Maisano, and Luca Mastrogiacomo. Ranking aggregation techniques. In *Rankings and Decisions in Engineering*, pages 85–160. Springer, 2022.

[Gra61] Franklin A. Graybill. An introduction to linear statistical models. pages 114–116, 1961.

[GSK69] James E. Grizzle, C. Frank Starmer, and Gary G. Koch. Analysis of categorical data by linear models. *Biometrics*, 25(3):489–504, 1969.

[GZ64] A. J. Goldman and Marvin Zelen. Weak generalized inverses and minimum variance linear unbiased estimation. *Journal of Research of the National Bureau of Standards Section B Mathematics and Mathematical Physics*, page 151, 1964.

[HOS20] Julien Hendrickx, Alex Olshevsky, and Venkatesh Saligrama. Minimax rate for learning from pairwise comparisons in the BTL model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4193–4202. PMLR, 13–18 Jul 2020.

[Hun04] David R. Hunter. Mm algorithms for generalized bradley-terry models. *The Annals of Statistics*, 32(1):384–406, 2004. URL `http://www.jstor.org/stable/3448514`.

[KT21] Eglantine Karlé and Hemant Tyagi. Dynamic ranking with the btl model: A nearest neighbor based rank centrality method. *arXiv preprint arXiv:2109.13743*, 2021.

[Lin10] Shili Lin. Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):555–570, 2010.

[Luc59] Robert D Luce. *Individual Choice Behavior: A theoretical analysis, New York, NY: John Willey and Sons.* Inc, 1959.

[Mal57] C. L. Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.

[NOS12] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Iterative ranking from pair-wise comparisons. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[PB60] Robert N Pendergrass and R Bradley. Ranking in triple comparisons. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, and H. B. Mann, editors, *Contributions to probability and statistics: Essays in honor of Harold Hotelling*, pages 331–351. Stanford University Press, Palo Alto, CA, 1960.

[Pen57] Robert Nixon Pendergrass. *The rank analysis of triple comparisons.* PhD thesis, Virginia Polytechnic Institute, 1957.

[Pla75] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975.

[RK67] P. V. Rao and L. L. Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967.

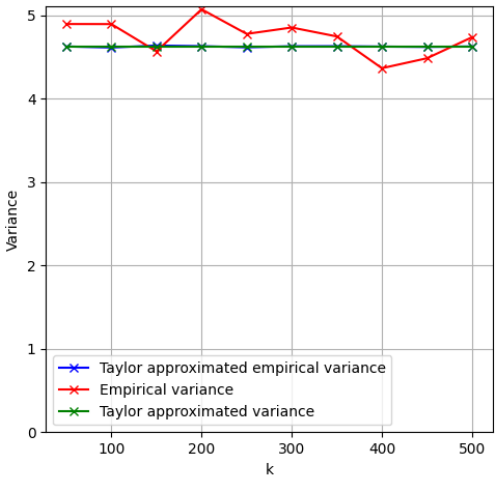
[SW17] Nihar B Shah and Martin J Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *The Journal of Machine Learning Research*, 18(1):7246–7283, 2017.

[Thu27] L. L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. doi:10.1037/h0070288.

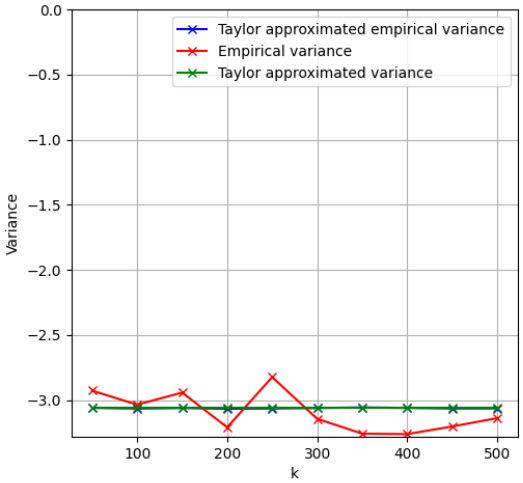
[Win21] Maxime Winand. Learning from pairwise comparisons : an empirical analysis. *Ecole polytechnique de Louvain, Université catholique de Louvain*, 2021. Prom.: Hendrickx, Julien.

Appendices

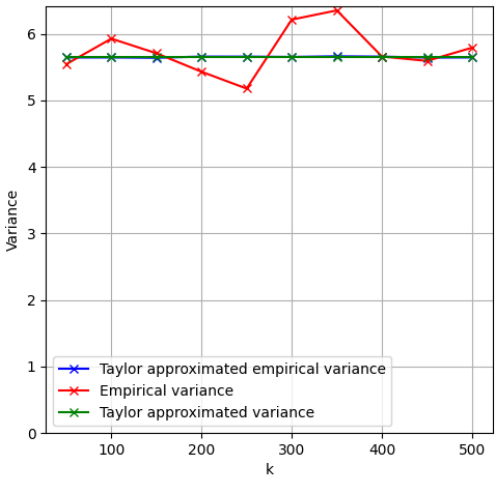
Appendix A
Empirical results for the method designed



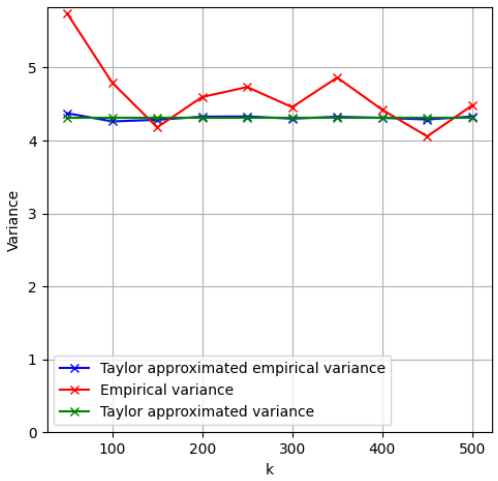
(a) Variance Rao-Kupper: $\mathbb{V}\left[\frac{h_{i,j}}{1-h_{i,j}}\right]$



(b) Covariance Rao-Kupper: $\mathbb{C}\left[\frac{h_{i,j}}{1-h_{i,j}}, \frac{h_{j,i}}{1-h_{j,i}}\right]$

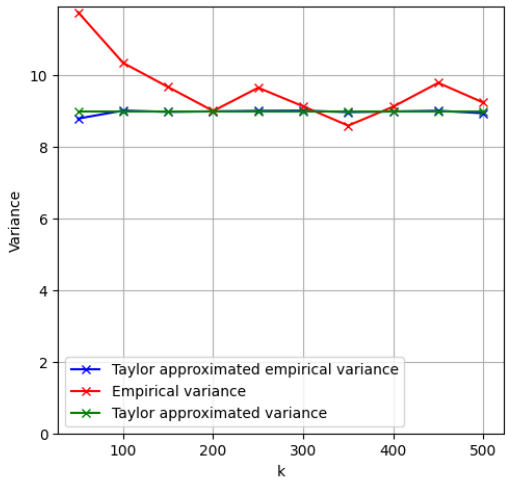


(c) Variance Davidson: $\mathbb{V}\left[\frac{h_{i,j}}{h_{j,i}}\right]$

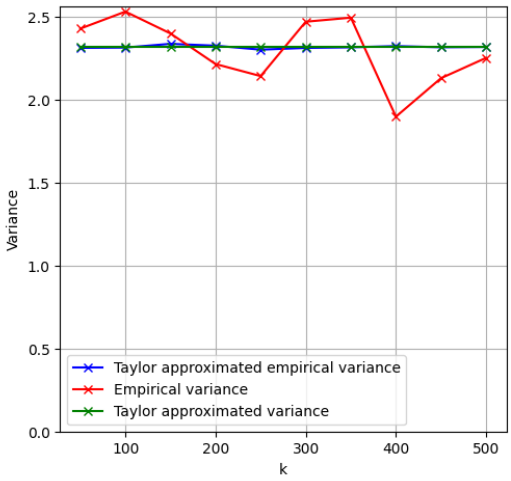


(d) Covariance Davidson: $\mathbb{C}\left[\frac{h_{j,i}}{h_{i,j}}, \frac{h_{i,j}}{h_{i=j}}\right]$

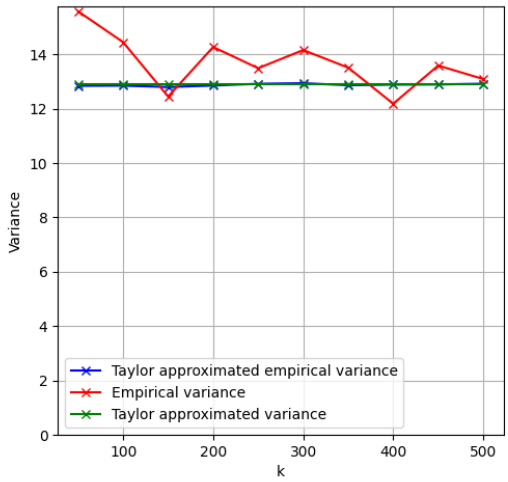
Figure A.1: Variance verification for pairwise comparison models with ties (Experiment done on 2 items with scores $\boldsymbol{w} \in [1, 20]$ drawn over 1000 simulations).



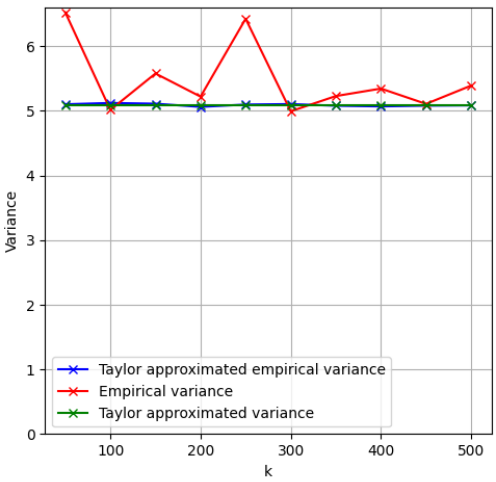
(a) Variance Plackett-Luce: $\mathbb{V}\left[\frac{h_{i,j}}{h_{j,i}}\right]$



(b) Covariance Plackett-Luce: $\mathbb{C}\left[\frac{h_{i,j}}{h_{j,i}}, \frac{h_{i,k}}{h_{k,i}}\right]$



(c) Variance MBT: $\mathbb{V}\left[\frac{h_{i,j,k}}{h_{j,k,i}}\right]$



(d) Covariance MBT: $\mathbb{C}\left[\frac{h_{i,k,j}}{h_{j,k,i}}, \frac{h_{i,k,j}}{h_{i,j,k}}\right]$

Figure A.2: Variance verification for triple comparison models (Experiment done on 3 items with scores $\boldsymbol{w} \in [1, 20]$ drawn over 1000 simulations).

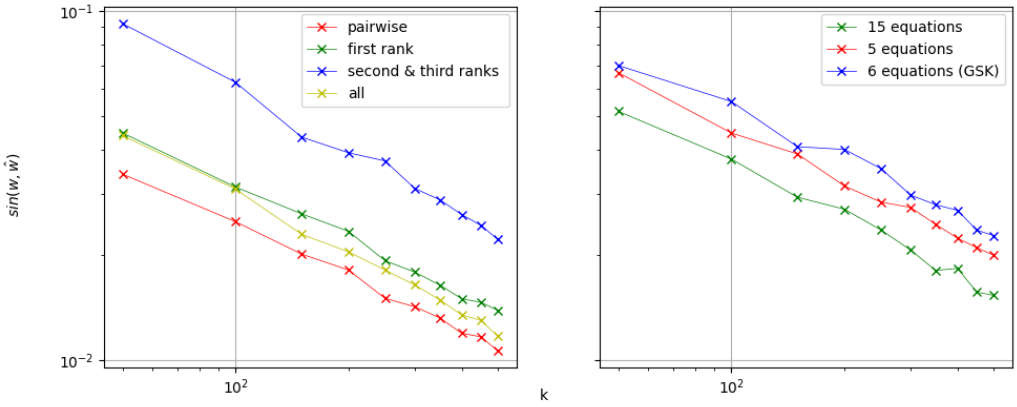


Figure A.3: Equation choice for the triple comparison models. Plackett-Luce and Mallows-Bradley-Terry models are given on the left and right figures respectively. The results are obtained for a set of $N = 10$ items on a complete graph (all possible triplets of items are compared to each other's) and the scores $\boldsymbol{w} \in [1, 20]$ are drawn over 100 simulations.

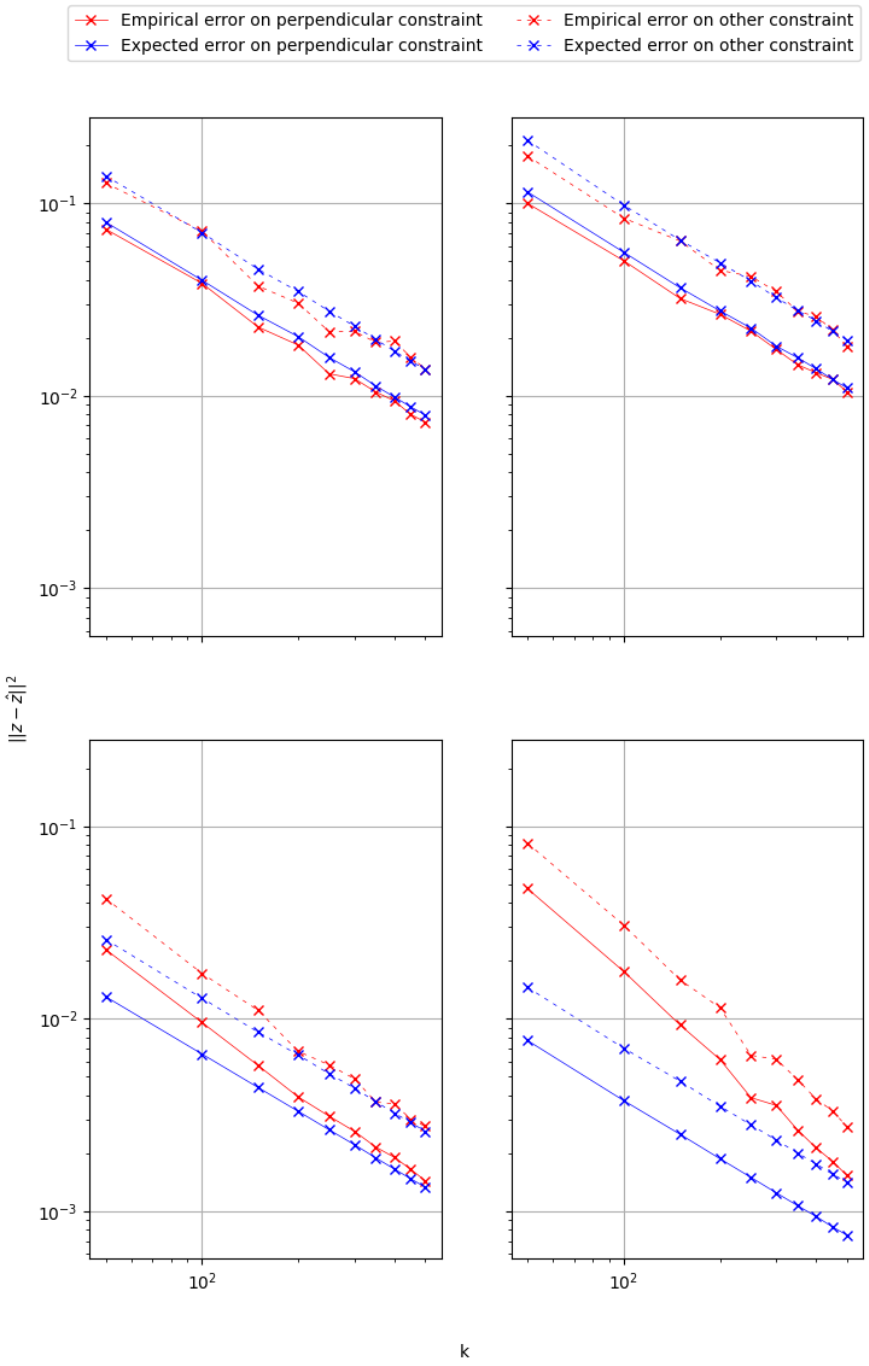


Figure A.4: Illustration of the expectation and empirical result of the Euclidean norm between a vector of parameter $\boldsymbol{z}$ and its estimator $\hat{\boldsymbol{z}}$ found using our weighted least squares method. This has been computed for the orthogonal constraint $\sum_{i=1}^{N} \log w_i = 1$ with $\boldsymbol{z} = \left[\log(\boldsymbol{\theta})' \quad \log(\boldsymbol{w})'\right]$ (full lines) and for a another constraint (dashed lines). This has been performed for the 4 models studied (Rao-Kupper ($\theta = 1.5$), Davidson ($\theta = 0.5$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right).

Appendix B
Complementary information about the computation of the variance

The expression 4.3:

$$\mathbb{C}\left[\log\frac{h_1}{h_2},\log\frac{h_3}{h_4}\right]\approx\frac{1}{p_1p_3}\mathbb{C}\left[h_1,h_3\right]-\frac{1}{p_1p_4}\mathbb{C}\left[h_1,h_4\right]-\frac{1}{p_2p_3}\mathbb{C}\left[h_2,h_3\right]+\frac{1}{p_2p_4}\mathbb{C}\left[h_2,h_4\right]$$

will be developed in detail in this appendix for the two triple comparisons models.

B.1 Mallow-Bradley-Terry

Let α, β, γ and δ be some ranking of the items i,j,k. Using the above expression and the fact that h_α follows a multinomial distribution with a number of experiments equal to $k_{\boldsymbol{ijk}}^{-1}$ we obtain in the different considered cases,

- Four different triplets:

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta},\log\frac{h_\gamma}{h_\lambda}\right]\approx\frac{1}{k_{\boldsymbol{ijk}}}\frac{p_\beta p_\lambda}{p_\alpha p_\gamma}\left(-\frac{p_\alpha p_\gamma}{p_\beta p_\lambda}+\frac{p_\gamma p_\alpha p_\lambda}{p_\beta p_\lambda^2}+\frac{p_\alpha p_\beta p_\gamma}{p_\beta^2 p_\lambda}-\frac{p_\alpha p_\gamma p_\beta p_\lambda}{p_\beta^2 p_\lambda^2}\right)$$

- Equal numerators:

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta},\log\frac{h_\alpha}{h_\lambda}\right]\approx\frac{1}{k_{\boldsymbol{ijk}}}\frac{p_\beta p_\lambda}{p_\alpha^2}\left(\frac{p_\alpha(1-p_\alpha)}{p_\beta p_\lambda}+\frac{p_\alpha^2 p_\lambda}{p_\beta p_\lambda^2}+\frac{p_\alpha^2 p_\beta}{p_\beta^2 p_\lambda}-\frac{p_\alpha^2 p_\beta p_\lambda}{p_\beta^2 p_\lambda^2}\right)$$

- Equal denominators:

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta},\log\frac{h_\gamma}{h_\beta}\right]\approx\frac{1}{k_{\boldsymbol{ijk}}}\frac{p_\beta^2}{p_\alpha p_\gamma}\left(-\frac{p_\alpha p_\gamma}{p_\beta^2}+\frac{p_\alpha p_\gamma p_\beta}{p_\beta^3}+\frac{p_\alpha p_\gamma p_\beta}{p_\beta^3}+\frac{p_\alpha p_\gamma p_\beta(1-p_\beta)}{p_\beta^4}\right)$$

- One numerator equal the denominator of the other term:

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta},\log\frac{h_\gamma}{h_\alpha}\right]\approx\frac{1}{k_{\boldsymbol{ijk}}}\frac{p_\beta p_\alpha}{p_\alpha p_\gamma}\left(-\frac{p_\alpha p_\gamma}{p_\beta p_\alpha}-\frac{p_\alpha(1-p_\alpha)p_\gamma}{p_\beta p_\alpha^2}+\frac{p_\alpha p_\beta p_\gamma}{p_\beta^2 p_\alpha}-\frac{p_\alpha^2 p_\gamma p_\beta}{p_\beta^2 p_\alpha^2}\right)$$

- Equal numerators and denominators:

$$\begin{aligned}\mathbb{C}\left[\log\frac{h_\alpha}{h_\beta},\log\frac{h_\alpha}{h_\beta}\right]&=\mathbb{V}\left[\log\frac{h_\alpha}{h_\beta}\right]\\&\approx\frac{1}{k_{\boldsymbol{ijk}}}\frac{p_\beta^2}{p_\alpha^2}\left(\frac{p_\alpha(1-p_\alpha)}{p_\beta^2}+\frac{p_\alpha^2 p_\beta}{p_\beta^3}+\frac{p_\alpha^2 p_\beta}{p_\beta^3}+\frac{p_\alpha^2 p_\beta(1-p_\beta)}{p_\beta^4}\right)\end{aligned}$$

Developing the above expressions allow to obtain the final covariances given by (4.12).

B.2 Plackett-Luce

Let α and β represent the relative order of two items of the triplet i, j, k. Denote by α^{-1} and β^{-1} the opposite relative order of the two items. In order to compute the above expression, one must be able to evaluate $\mathbb{C}\left[h_\alpha, h_\beta\right]$ for all the possibilities regarding α and β. Those computations are done below using first properties of the covariance and then the fact that h_* for any ranking $*$ of the items i, j, k follows a multinomial distribution with a number of experiments equal to $k_{\boldsymbol{ijk}}$ as h_* represents the fraction of times the ranking $*$ has been observed.

$$
\begin{aligned}
\mathbb{C}\left[h_{i,j}, h_{i,k}\right] &= \mathbb{C}\left[h_{i,j,k} + h_{i,k,j} + h_{k,i,j}, h_{i,j,k} + h_{i,k,j} + h_{j,i,k}\right] \\
&= \mathbb{C}\left[h_{i,j,k}, h_{i,j,k}\right] + \mathbb{C}\left[h_{i,j,k}, h_{i,k,j}\right] + \mathbb{C}\left[h_{i,j,k}, h_{j,i,k}\right] \\
&+ \mathbb{C}\left[h_{i,k,j}, h_{i,j,k}\right] + \mathbb{C}\left[h_{i,k,j}, h_{i,k,j}\right] + \mathbb{C}\left[h_{i,k,j}, h_{j,i,k}\right] \\
&+ \mathbb{C}\left[h_{k,i,j}, h_{i,j,k}\right] + \mathbb{C}\left[h_{k,i,j}, h_{i,k,j}\right] + \mathbb{C}\left[h_{k,i,j}, h_{j,i,k}\right] \\
&= \frac{1}{k_{\boldsymbol{ijk}}}\Big(p_{i,j,k}(1 - p_{i,j,k}) - p_{i,j,k}p_{i,k,j} - p_{i,j,k}p_{j,i,k} \\
&\qquad - p_{i,k,j}p_{i,j,k} + p_{i,k,j}(1 - p_{i,k,j}) - p_{i,k,j}p_{j,i,k} \\
&\qquad - p_{k,i,j}p_{i,j,k} - p_{k,i,j}p_{i,k,j} - p_{k,i,j}p_{j,i,k}\Big) \\
&= \frac{1}{k_{\boldsymbol{ijk}}}\Big(p_{i,j,k}(1 - p_{i,k}) + p_{i,k,j}(1 - p_{i,k}) - p_{k,i,j}p_{i,k}\Big) \\
&= \frac{1}{k_{\boldsymbol{ijk}}}\Big(p_{i,j,k} + p_{i,k,j} - p_{i,j}p_{i,k}\Big)
\end{aligned}
$$

The other correlations are obtain in the same way and are used to find the estimation of $\mathbb{C}\left[\log\frac{h_\alpha}{h_{\alpha^{-1}}}, \log\frac{h_\beta}{h_{\beta^{-1}}}\right]$ for the different possible combinations of α and β:

- $\alpha = i, j$ and $\beta = i, k$

$$
\begin{aligned}
\mathbb{C}\left[\log\frac{h_{i,j}}{h_{j,i}}, \log\frac{h_{i,k}}{h_{k,i}}\right] \approx &\frac{1}{p_{i,j}p_{i,k}}\frac{1}{k_{\boldsymbol{ijk}}}\Bigg((p_{i,j,k} + p_{i,k,j} - p_{i,j}p_{i,k}) - \frac{p_{i,k}}{p_{k,i}}(-p_{i,j}p_{k,i} + p_{k,i,j}) \\
&- \frac{p_{i,j}}{p_{j,i}}(-p_{j,i}p_{i,k} + p_{j,i,k}) + \frac{p_{i,j}p_{i,k}}{p_{j,i}p_{k,i}}(-p_{j,i}p_{k,i} + p_{j,k,i} + p_{k,j,i})\Bigg) \\
= &\frac{1}{p_{i,j}p_{i,k}}\frac{1}{k_{\boldsymbol{ijk}}}\left(\frac{p_{j,k,i} + p_{k,j,i}}{p_{k,i}p_{j,i}} - 1\right)
\end{aligned}
$$

- $\alpha = i, j$ and $\beta = j, k$

$$
\begin{aligned}
\mathbb{C}\left[\log\frac{h_{i,j}}{h_{j,i}}, \log\frac{h_{j,k}}{h_{k,j}}\right] \approx &\frac{1}{p_{i,j}p_{j,k}}\frac{1}{k_{\boldsymbol{ijk}}}\Bigg((p_{i,j,k} - p_{i,j}p_{j,k}) - \frac{p_{j,k}}{p_{k,j}}(-p_{i,j}p_{k,j} + p_{k,i,j} + p_{i,k,j}) \\
&- \frac{p_{i,j}}{p_{j,i}}(-p_{j,i}p_{j,k} + p_{j,i,k} + p_{j,k,i}) + \frac{p_{i,j}p_{j,k}}{p_{j,i}p_{k,j}}(-p_{j,i}p_{k,j} + p_{k,j,i})\Bigg) \\
= &\frac{1}{p_{i,j}p_{j,k}}\frac{1}{k_{\boldsymbol{ijk}}}\left(\frac{p_{k,j,i}}{p_{k,j}p_{j,i}} - 1\right)
\end{aligned}
$$

- $\alpha = i, k$ and $\beta = j, k$

$$\begin{aligned}\mathbb{C}\left[\log\frac{h_{i,k}}{h_{k,i}}, \log\frac{h_{j,k}}{h_{k,j}}\right] \approx & \frac{1}{p_{i,k}p_{j,k}}\frac{1}{k_{\boldsymbol{ijk}}}\left((p_{i,j,k} + p_{j,i,k} - p_{i,k}p_{j,k}) - \frac{p_{j,k}}{p_{k,j}}(-p_{i,k}p_{k,j} + p_{i,k,j})\right. \\ & \left. - \frac{p_{i,k}}{p_{k,i}}(-p_{k,i}p_{j,k} + p_{j,k,i}) + \frac{p_{i,k}p_{j,k}}{p_{k,i}p_{k,j}}(-p_{k,i}p_{k,j} + p_{k,i,j} + p_{k,j,i})\right) \\ = & \frac{1}{p_{i,k}p_{j,k}}\frac{1}{k_{\boldsymbol{ijk}}}\left(\frac{p_{k,i,j} + p_{k,j,i}}{p_{k,j}p_{k,i}} - 1\right)\end{aligned}$$

- $\alpha = \beta$

$$\mathbb{C}\left[\log\frac{h_\alpha}{h_{\alpha^{-1}}}, \log\frac{h_\alpha}{h_{\alpha^{-1}}}\right] = \mathbb{V}\left[\log\frac{h_\alpha}{h_{\alpha^{-1}}}\right] \approx \frac{1}{k_{\boldsymbol{ijk}}}\left(\frac{1}{p_\alpha} + \frac{1}{p_{\alpha^{-1}}}\right)$$

Those expressions are summarized in Chapter 4 in (4.14).

Appendix C
Details about Chapter 7

C.1 Possibly unobservable parameters β

Remember that when the matrix X' of problem 5.1 has a rank $p' \leq p$, the parameters β are unobservable and an alternative problem was proposed in [GZ64] given by Theorem 5.2.1 to overcome the issue. A unique estimator of β can be obtained from (5.3) by adding $p - p'$ conditions linearly independents of the rows of X' on β, i.e. satisfying $S'\beta = m$ with $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. This estimator will be the BLUE of the new parameter vector β_c such that:

$$Y = X'\beta_c + \epsilon \quad m = S'\beta_c \quad \textit{with } \text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p,\ \mathbb{E}(Y) = X'\beta_c,\ \mathbb{V}(Y) = \sigma^2 I$$

Theorem C.1.1. Consider the system (5.1) with $\mathbb{V}[Y] = \mathbb{V}[\epsilon] = \sigma^2 I$. Among the possible vectors β, denote β_c as the unique one satisfying the additional conditions $S'\beta = m$ for any matrix S of size $p \times (p - p')$ such that $\text{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. The best linear unbiased estimator of β_c from this redefined system exists and is given by

$$\hat{\beta}_c = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix} + B(S'B)^{-1}m$$

for any matrix L' of size $p' \times p$ and rank p' and the corresponding $\hat{\gamma}$ from Theorem 5.2.1 and where B is a basis of $\text{Ker}(X')$.

Proof. Let us first restrict ourselves to homogeneous constraints of the form $S'\beta = 0$. Theorem 5.2.1 states that $\hat{\gamma} = L'\hat{\beta}$ is the BLUE estimator of $\gamma = L'\beta$. Since $\hat{\gamma}$ is linear, i.e. $\hat{\gamma} = D'Y$, the following holds:

$$\hat{\beta}_c = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \hat{\gamma} \\ 0_{(p'-p)} \end{bmatrix} = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} D'Y \\ 0_{(p'-p)} \end{bmatrix} = \begin{bmatrix} L' \\ S' \end{bmatrix}_{p'}^{-1} D'Y$$

where $\left(\begin{bmatrix} L & S \end{bmatrix}'\right)_{p'}^{-1}$ is the matrix containing the first p' columns of the matrix. The estimator $\hat{\beta}_c$ is thus linear. As $\hat{\gamma}$ is unbiased, so is $\hat{\beta}_c$:

$$\begin{aligned} \mathbb{E}[\hat{\beta}_c] &= \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \mathbb{E}[\hat{\gamma}] \\ 0_{(p'-p)} \end{bmatrix} = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \gamma \\ 0_{(p'-p)} \end{bmatrix} \\ &= \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} L'\beta_c \\ 0_{(p'-p)} \end{bmatrix} = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} L' \\ S' \end{bmatrix} \beta_c = \beta_c \end{aligned}$$

where the fourth equality stems from the fact that $S'\beta_c = 0$.

Finally, $\hat{\beta}_c$ is also the best estimator among the unbiased linear estimators. For all vectors $\tilde{\beta}_c$ belonging to C, there is a vector

$$\begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix} = \begin{bmatrix} L' \\ S' \end{bmatrix} \tilde{\beta}_c \quad \text{such that} \quad \tilde{\beta}_c = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix}$$

From Theorem 5.2.1, it is known that $\hat{\gamma}$ is the BLUE estimator of $\gamma = L'\beta_c$, i.e.

$$(\mathbb{V}(\tilde{\gamma}) - \mathbb{V}(\hat{\gamma})) \succeq 0$$

Consequently,

$$\mathbb{V}\left(\begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) - \mathbb{V}\left(\begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) = \begin{bmatrix} \mathbb{V}(\tilde{\gamma}) - \mathbb{V}(\hat{\gamma}) & 0_{p' \times (p-p')} \\ 0_{(p-p') \times p'} & 0_{(p-p') \times (p-p')} \end{bmatrix} \succeq 0 \tag{C.1}$$

Using this vector, it is also true that

$$\begin{aligned} \mathbb{V}(\tilde{\beta}_c) - \mathbb{V}(\hat{\beta}_c) &= \mathbb{V}\left(\begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) - \mathbb{V}\left(\begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) \\ &= \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \left(\mathbb{V}\begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) \begin{bmatrix} L & S \end{bmatrix}^{-1} \end{aligned}$$

To be positive semi-definite, the following relation must be satisfied:

$$\forall x \in \mathbb{R}^p,\ x' \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \left(\mathbb{V}\begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) \begin{bmatrix} L & S \end{bmatrix}^{-1} x \geq 0$$

which can be reformulated as

$$\forall \tilde{y} \in \mathbb{R}^p \text{ such that } \tilde{y} = \begin{bmatrix} L & S \end{bmatrix}^{-1} x,\ \tilde{y}' \left(\mathbb{V}\begin{bmatrix} \tilde{\gamma} \\ 0_{(p-p')} \end{bmatrix} - \mathbb{V}\begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix}\right) \tilde{y} \geq 0$$

Using the relation (C.1), this relation is always satisfied.

Consider now any linear constraints of the form $S'\beta_c = m$. Because β_c belongs to a class of equivalence satisfying $Y = X'\beta + \epsilon$, we know that the vector is defined by $\beta_c = \beta_0 + v_{X'}$ where β_0 is an element of that class and $v_{X'} \in \text{Ker}(X')$. Define β_0 as the vector satisfying $Y = X'\beta_0 + \epsilon$ and $S'\beta_0 = 0$. Then the constraints are such that $S'\beta_c = S'(\beta_0 + v_{X'}) = S'v_{X'} = m$. Denote by B a basis of $\text{Ker}(X')$, then $v_{X'} = \sum_{i=1}^{p-p'} B_i\alpha_i = B\alpha$. The constraints becomes $S'B\alpha = m$ and because S' and B are full rank matrices, we obtain $\alpha = (S'B)^{-1}m$. In other words, the vector one wants to approximate can be seen as $\beta_c = \beta_0 + B(S'B)^{-1}m$.

The BLUE of β_c is $\hat{\beta}_c = \hat{\beta}_0 + B(S'B)^{-1}m$ where $\hat{\beta}_0$ is the BLUE of β_0 which is given by $\hat{\beta}_0 = (Q'QXX'Q'Q)^{\dagger} XY$ as proven above. We indeed have that $\hat{\beta}_c$

is a linear and unbiased estimator of β_c since it is a linear combination of the observations to which a constant vector $v_{X'}$ is added and it holds that

$$\mathbb{E}[\hat{\beta}_c] = \mathbb{E}[\hat{\beta}_0 + B(S'B)^{-1}m] = \beta_0 + B(S'B)^{-1}m = \beta_c$$

since $\mathbb{E}[\hat{\beta}_0] = \beta_0$ as $\hat{\beta}_0$ is an unbiased estimator of β_0.

We still have to prove that the found estimator is the best one in the sense that it minimizes the variance. Take $\tilde{\beta}_c$, another unbiased linear estimator of β_c. Then $\tilde{\beta}_c$ can be written as $\tilde{\beta}_c = \tilde{\beta}_0 + \tilde{v}_{X'}$ where $\tilde{\beta}_0$ is a linear unbiased estimator of β_0 and $\tilde{v}_{X'} \in \mathrm{Ker}(X')$. Because $\hat{\beta}_0$ is the BLUE of β_0 it holds that

$$\mathbb{V}(\hat{\beta}_c) = \mathbb{V}(\hat{\beta}_0 + v_{X'}) = \mathbb{V}(\hat{\beta}_0) \preceq \mathbb{V}(\tilde{\beta}_0) = \mathbb{V}(\tilde{\beta}_c)$$

□

C.2 Possibly non independent and non-homoscedastic observations

In this case, the variance of the estimator is not the identity matrix anymore and it might be a singular matrix. Considering Theorem 5.3.2, one may be interested to make the estimator of β unique by adding $p - p'$ conditions linearly independents of the rows of X' on β, i.e. satisfying $S'\beta = m$ with $\mathrm{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. This estimator will be the BLUE of the new parameter vector β_c such that:

$$Y = X'\beta_c + \epsilon \quad m = S'\beta_c \quad \textit{with } \mathrm{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p,\ \mathbb{E}(Y) = X'\beta_c,\ \mathbb{V}(Y) = V$$

Theorem C.2.1. Consider the system (5.1) with X' possibly of non-full rank, i.e. $\mathrm{rank}(X') = p' \leq p$, and $\mathbb{V}[\epsilon] = V$ possibly singular, i.e. $\mathrm{rank}(V) = m' \leq m$. Among the possible vectors β, denote β_c as the unique one satisfying the additional conditions $S'\beta = m$ for any matrix S of size $p \times (p - p')$ such that $\mathrm{rank}\left(\begin{bmatrix} X & S \end{bmatrix}\right) = p$. If $\mathrm{Ker}(V) \subseteq \mathrm{Ker}(X)$, the best linear unbiased estimator of β_c from this redefined system exists and is for any matrix L' of size $p' \times p$ and rank p' and the corresponding $\hat{\gamma}$ from Theorem 5.3.2:

$$\hat{\beta}_c = \begin{bmatrix} L' \\ S' \end{bmatrix}^{-1} \begin{bmatrix} \hat{\gamma} \\ 0_{(p-p')} \end{bmatrix} + B(S'B)^{-1}m$$

where B is a basis of $\mathrm{Ker}(X')$.

Proof. By Theorem 5.3.1, the problem can be reformulated as

$$\tilde{Y} = \tilde{X}'\beta + \tilde{\epsilon} \quad \text{such that} \quad K'\beta = m$$

where $\tilde{Y} = T'Y$, $\tilde{X} = XT$, $\tilde{\epsilon} = T'\epsilon$ and T defined as in the Lemma.

Since $\mathbb{V}[\tilde{Y}] = \mathbb{V}[\tilde{\epsilon}] = I$, Theorem C.1.1 can be used with $\tilde{Y}$, $\tilde{X}$, $\tilde{\epsilon}$ instead of Y, X, ϵ and with $TT' = V^{\dagger}$.

□

Appendix D
Complementary information about the rank of *M'* for the models extending BTL

This appendix aims to convince the reader about the fact that the matrix M' of dimension $n_{eq} \times \tilde{N}$ in the problem of finding the BLUE of $\tilde{\boldsymbol{z}}$ in the noisy linear system $\log \boldsymbol{R} = M'\tilde{\boldsymbol{z}} + \boldsymbol{\epsilon}$ is of rank $\tilde{N} - 1$.

D.1 Pairwise comparison models with ties

In Chapter 6, it is stated that the matrix M' can be written as

$$M' = D \begin{bmatrix} 1 & 0 \\ 0 & B' \end{bmatrix}$$

where B' is the incidence matrix of the graph $G = (V, E)$ such that the vertices are all the items and E links to items together if they are compared together and D is a full rank matrix of size $n_{eq} \times (n_s + 1)$ where n_s is the number of pairs of items being compared to each other. This Section will provide the full rank matrix D for the two pairwise comparison models with ties.

D.1.1 Rao-Kupper model

From Chapter 6 we know that for the Rao-Kupper model

$$M' = \begin{bmatrix} -\mathbb{1} & B' \\ -\mathbb{1} & -B' \end{bmatrix}$$

On can thus take

$$D = \begin{bmatrix} -\mathbb{1}_{n_s} & I_{n_s} \\ -\mathbb{1}_{n_s} & -I_{n_s} \end{bmatrix}$$

To prove that this matrix is full rank, we will show that

$$\exists \alpha \in \mathbb{R}^{n_s+1} \text{ such that } \alpha' \tilde{D} = e_i \ \forall i \in \{1, \cdots, n_s + 1\}$$

where e_i is the null vector with entry i equal to 1 and $\tilde{D}$ is the matrix composed of the first $n_s + 1$ rows of D. This indeed proves that the rank of D is $n_s + 1$ as one can recover the canonical basis of $\mathbb{R}^{n_s+1}$ from the rows of D and since the matrix is of dimension $n_{eq} \times (n_s + 1)$ with $n_{eq} = 2n_s$, D is full rank. The above statement holds as

$$\alpha' \tilde{D} = \begin{bmatrix} -\sum_i \alpha_i & \alpha_1 - \alpha_{n_s+1} & \alpha_2 & \alpha_3 & \cdots & \alpha_{n_s} \end{bmatrix}$$

meaning that

$$e_1 = \left(-\frac{1}{2}e_1 - \frac{1}{2}e_{n_s+1}\right)\tilde{D}$$
$$e_i = \left(-\frac{1}{2}e_1 + e_{i-1} - \frac{1}{2}e_{n_s+1}\right)\tilde{D} \qquad \forall i \in \{2, \cdots, n_s+1\}$$

D.1.2 Davidson model

From Chapter 6 we know that for the Davidson model

$$M' = \begin{bmatrix} 0 & B' \\ \mathbb{1} & -\frac{1}{2}B' \\ \mathbb{1} & \frac{1}{2}B' \end{bmatrix}$$

On can thus take

$$D = \begin{bmatrix} 0_{n_s} & I_{n_s} \\ \mathbb{1}_{n_s} & -\frac{1}{2}I_{n_s} \\ \mathbb{1}_{n_s} & \frac{1}{2}I_{n_s} \end{bmatrix}$$

To prove that this matrix is full rank, as for the Rao-Kupper model, we will show that

$$\exists \alpha \in \mathbb{R}^{n_s+1} \text{ such that } \alpha'\tilde{D} = e_i \ \forall i \in \{1, \cdots, n_s+1\}$$

where e_i is the null vector with entry i equal to 1 and $\tilde{D}$ is the matrix composed of the first n_s+1 rows of D which is of dimension $n_{eq} \times (n_s+1)$ with $n_{eq} = 3n_s$. The above statement holds as

$$\alpha'\tilde{D} = \begin{bmatrix} \alpha_{n_s+1} & \alpha_1 - \frac{\alpha_{n_s+1}}{2} & \alpha_2 & \alpha_3 & \cdots & \alpha_{n_s} \end{bmatrix}$$

meaning that

$$e_1 = \left(\frac{1}{2}e_1 + e_{n_s+1}\right)\tilde{D}$$
$$e_i = e_{i-1}\tilde{D} \qquad \forall i \in \{2, \cdots, n_s+1\}$$

D.2 Triple comparison models

In Chapter 6, it is stated that the matrix M' can be written as

$$M' = DB'$$

where B' is the incidence matrix of the graph with the items as vertices and that link two vertices if they belong to at least one triplet of items being compared to one another and D is a full rank matrix of size $n_{eq} \times n_{edge}$ (where n_{edge} represents the number of edges composing the graph). This Section will provide the full rank matrix D for the two triple comparison models.

D.2.1 Plackett-Luce model

From Chapter 6, we know that M' is composed of rows equal to $e_i - e_j$ for two items i and j being part of a triplet of items compared to each other. Note that several rows of the matrix could be equal since two items can be part of several different triplet of items. The matrix D is then composed of rows equal to e_p. By the definition of B' and because we only consider connected graph structures, we will always have at least one row of D equal to e_p for $p \in \{1, \cdots, n_{edge}\}$ meaning that the matrix D contains the identity matrix and is thus full rank.

For a graph composed of four items (thus 4 triplets) we have

$$M' = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}$$

where the equations are sorted by triplets (the three first equations given by the first three lines of M' corresponds to the first triplet of items $1, 2, 3$ and so on for triplets $1, 2, 4$; $1, 3, 4$; $2, 3, 4$). One indeed observe that the matrix D is full rank since the rows $1, 2, 5, 3, 6, 9$ form the identity matrix.

D.2.2 Mallows-Bradley-Terry model

From Chapter 6, we know that a row of M' corresponding to the equation comparing ranking π and π' of items i, j, k has coefficients corresponding to commutation of indices in the two rankings since

$$\log\left(\frac{h_\pi}{h_{\pi'}}\right) \approx (\pi'^{-1} - \pi^{-1})(i)\log(w_i) + (\pi'^{-1} - \pi^{-1})(j)\log(w_j) + (\pi'^{-1} - \pi^{-1})(k)\log(w_k)$$

for all rankings $\pi, \pi' \in \{(i,j,k), (i,k,j), (j,i,k), (j,k,i), (k,i,j), (k,j,i)\}$ and all triplet of items (i, j, k).

For one triplet of item i, j, k we have thus

$$\tilde{M} = \begin{bmatrix} 0 & -1 & 1 \\ -1 & 1 & 0 \\ -2 & 1 & 1 \\ -1 & -1 & 2 \\ -2 & 0 & 2 \\ -1 & 2 & -1 \\ -2 & 2 & 0 \\ -1 & 0 & 1 \\ -2 & 1 & 1 \\ -1 & 0 & 1 \\ 0 & -2 & 2 \\ -1 & -1 & 2 \\ 1 & -2 & 1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & -1 & -1 \\ 0 & -2 & 0 \\ -1 & 0 & 1 \\ -2 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -2 \\ 0 & -1 & -1 \\ 1 & 0 & -1 \\ 0 & 0 & -1 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

if the equations are in the following order

$$\left\{ \frac{h_{i,k,j}}{h_{i,j,k}}, \frac{h_{j,i,k}}{h_{i,j,k}}, \frac{h_{j,k,i}}{h_{i,j,k}}, \frac{h_{k,i,j}}{h_{i,j,k}}, \frac{h_{k,j,i}}{h_{i,j,k}}, \frac{h_{j,i,k}}{h_{i,k,j}}, \frac{h_{j,k,i}}{h_{i,k,j}}, \right.$$
$$\left. \frac{h_{k,i,j}}{h_{i,k,j}}, \frac{h_{k,j,i}}{h_{i,k,j}}, \frac{h_{j,k,i}}{h_{j,i,k}}, \frac{h_{k,i,j}}{h_{j,i,k}}, \frac{h_{k,j,i}}{h_{j,i,k}}, \frac{h_{k,i,j}}{h_{j,k,i}}, \frac{h_{k,j,i}}{h_{j,k,i}}, \frac{h_{k,j,i}}{h_{k,i,j}} \right\}$$

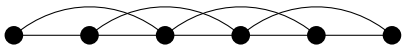
We observe that the matrix $\tilde{D}$ satisfying $\tilde{M} = \tilde{D}B'$ is full rank since the rows $2, 8, 1$ form the opposite of the identity matrix.

For a larger set of items, M and D will be composed of rows sets related to each triplet of items. For the set of rows associated to items i, j, k, the corresponding columns, i.e. columns i, j, k, form $\tilde{M}$ and $\tilde{D}$. Because $\tilde{D}$ is full rank and B' represents a connected graph, D will be full rank as well.

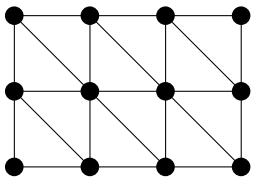
Appendix E
Empirical results for all the models studied

This Appendix contains the results presented in Chapter 7 for the different models extending the BTL model studied in this report, namely the Rao-Kupper, the Davidson, the Plackett-Luce and the Mallows-Bradley-Terry models.

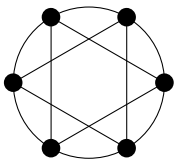
For the graph structures considered, the adaptation for triple comparisons models are given below (for the Erdös-Rényi graph structure, a triplet of items is connected with a probability $\log(N)/N$):



(a) Line structure



(b) 2D-grid structure



(c) Circle structure

Figure E.1: Adaptation of the different graph structures for triple comparison models.

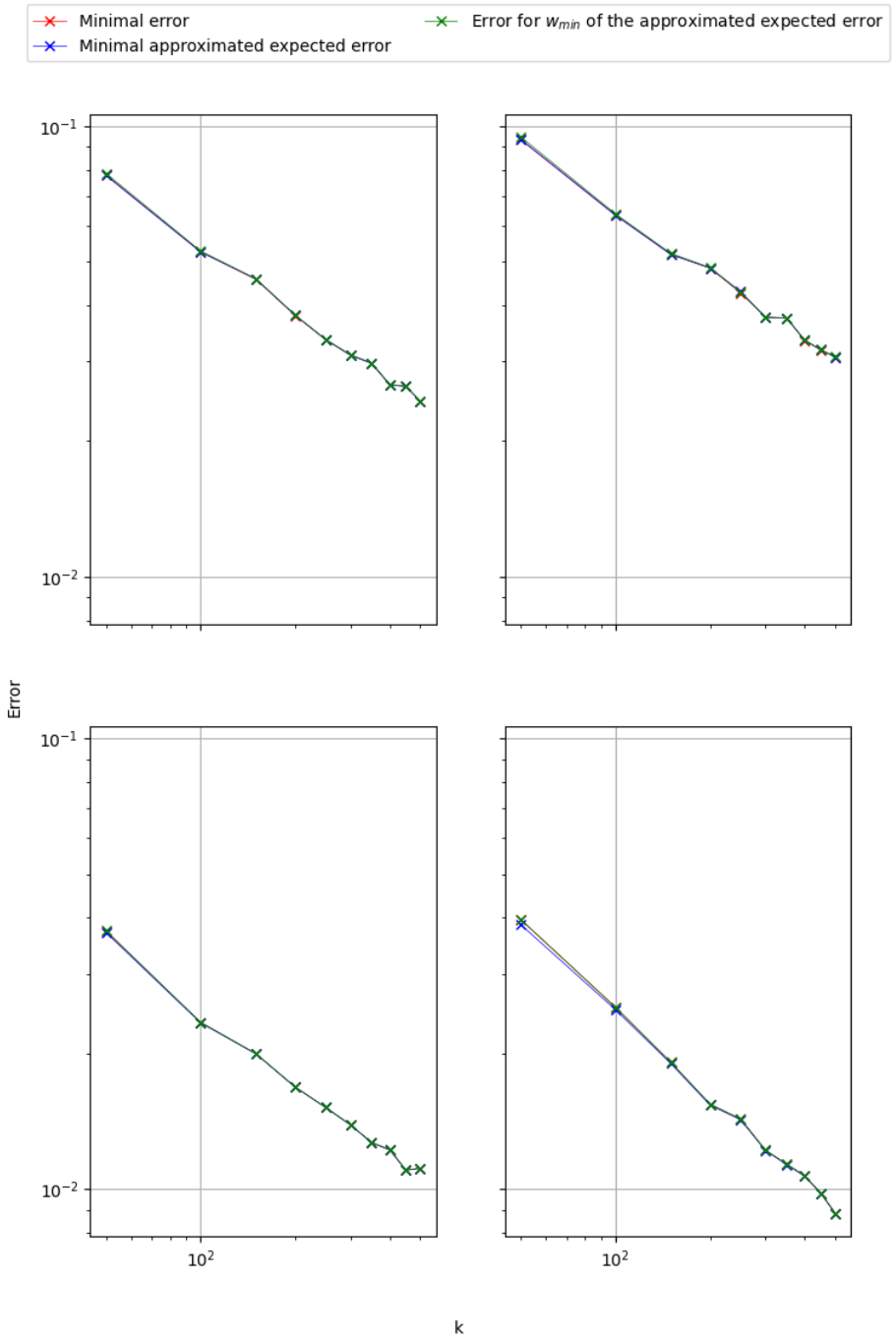


Figure E.2: Illustration of the error measures of the item weights for the different models (Rao-Kupper ($\theta = 1.5$), Davidson ($\theta = 0.5$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right). The red curve shows the sine error (minimal error obtained between the true weights $\boldsymbol{w}$ and it estimation $\hat{\boldsymbol{w}}_{min}$ belonging to the set $MV^{\dagger}M\hat{\boldsymbol{z}} = MV^{\dagger}\log \boldsymbol{R}$); the blue curve shows the error between the vectors $\log \boldsymbol{w}$ and $\log \hat{\boldsymbol{w}}$ for an estimator minimizing the weighted $2-$norm with weights D_w; the green curve shows the error between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$. The experiments have been done on a complete graph of 10 items with a constant number of comparisons k between comparison sets; scores are generated on 100 simulation such that $\boldsymbol{w} \in [1, 20]$

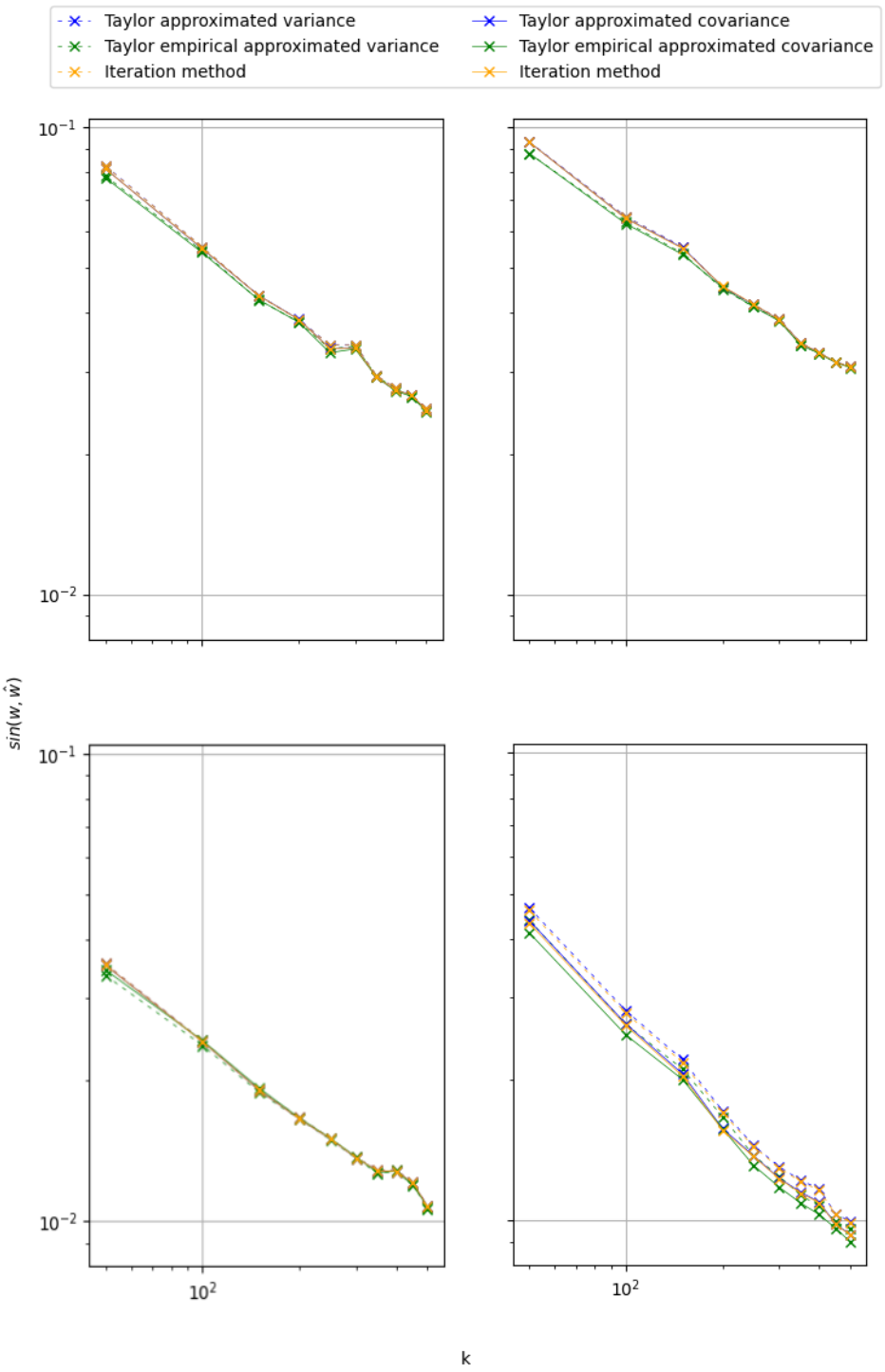


Figure E.3: Comparison of the performances obtained for the weighted least squares method with different weighted matrices for the different models (Rao-Kupper ($\theta = 1.5$), Davidson ($\theta = 0.5$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right).

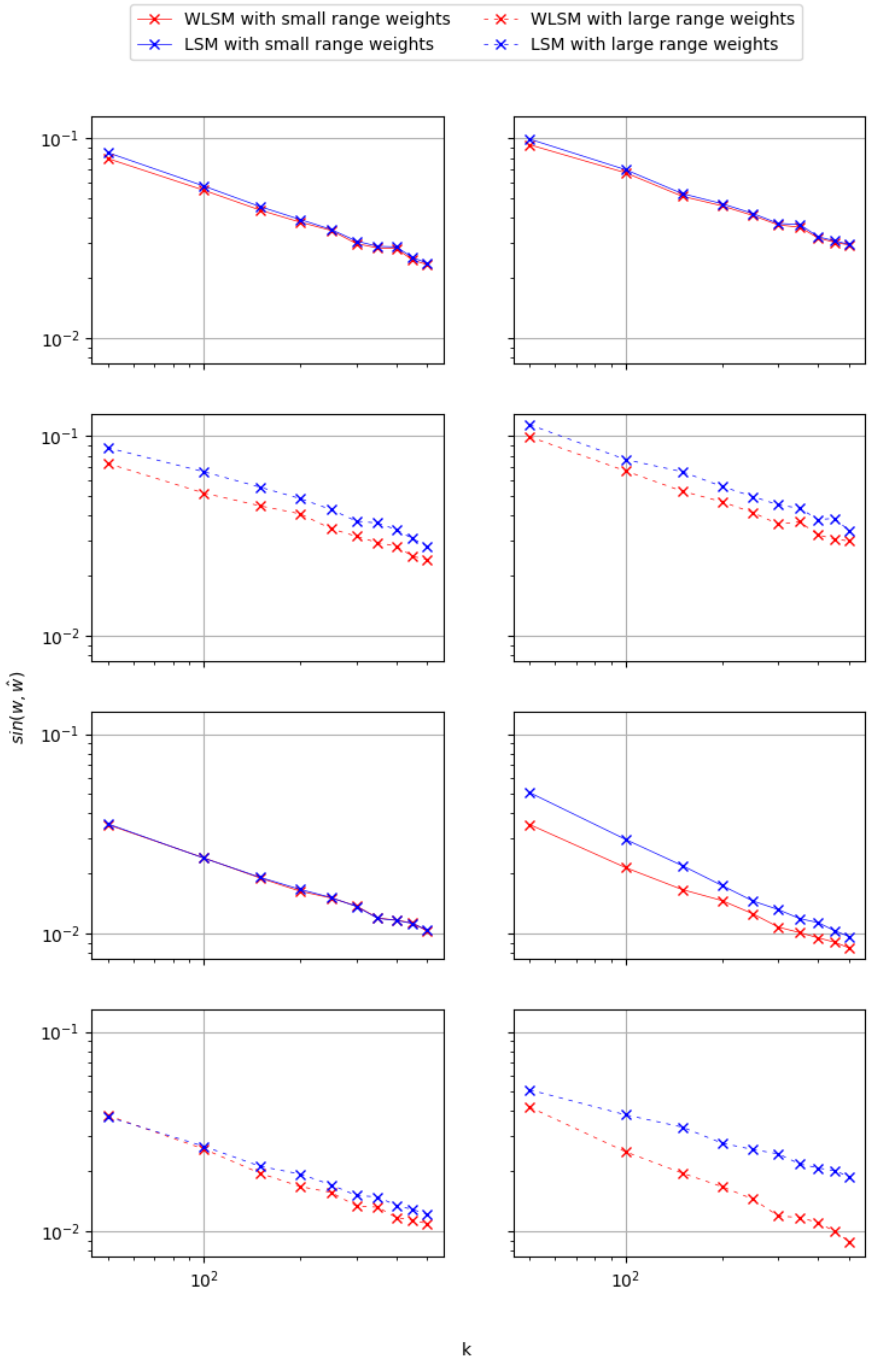


Figure E.4: Illustration of the impact of the variance in the weighted least squares method for the different models (Rao-Kupper ($\theta = 2$), Davidson ($\theta = 1$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right with two stacks figure representing one model).

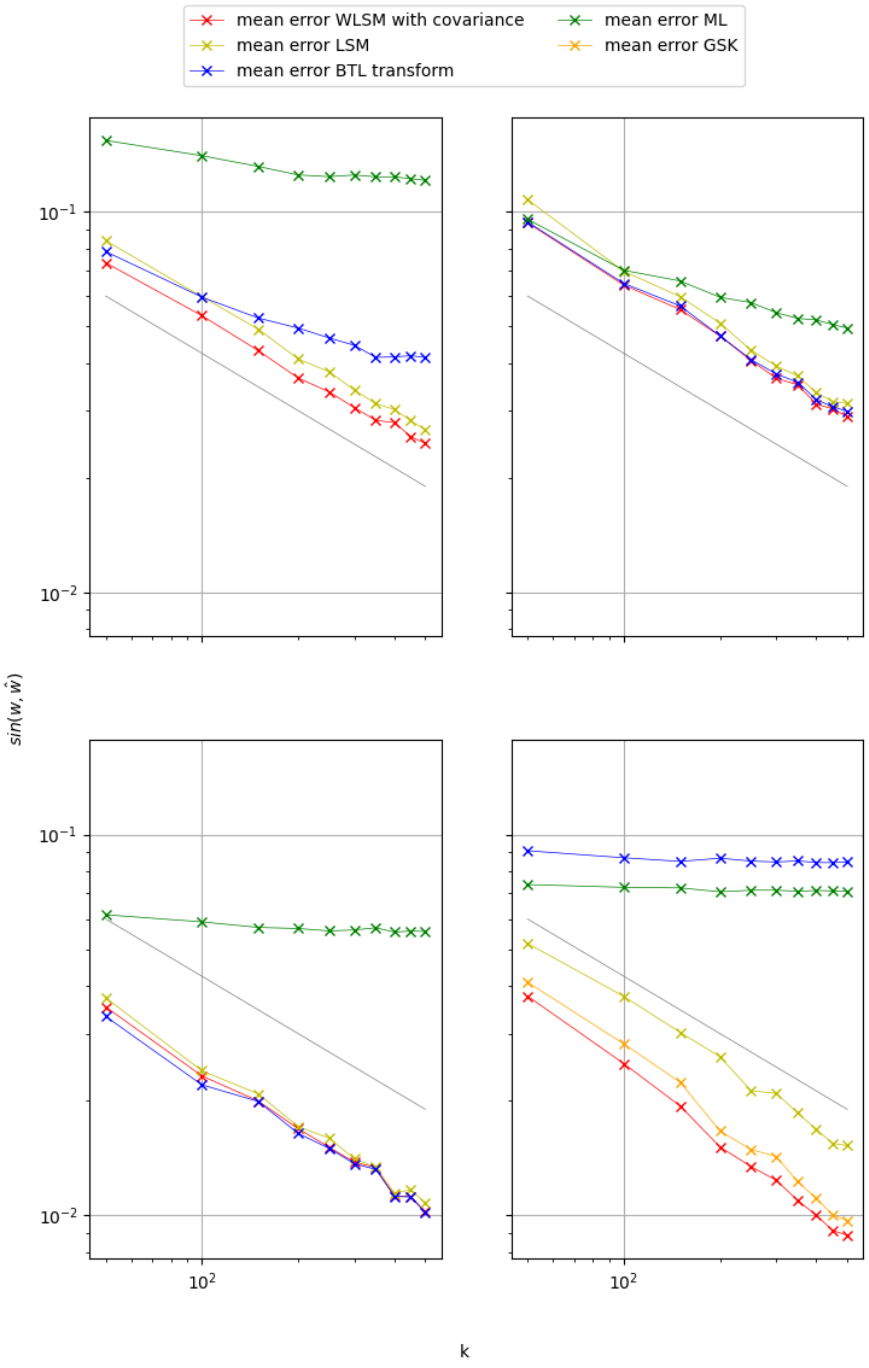


Figure E.5: Comparison of the different methods for the different models (Rao-Kupper ($\theta = 2$), Davidson ($\theta = 1$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right).

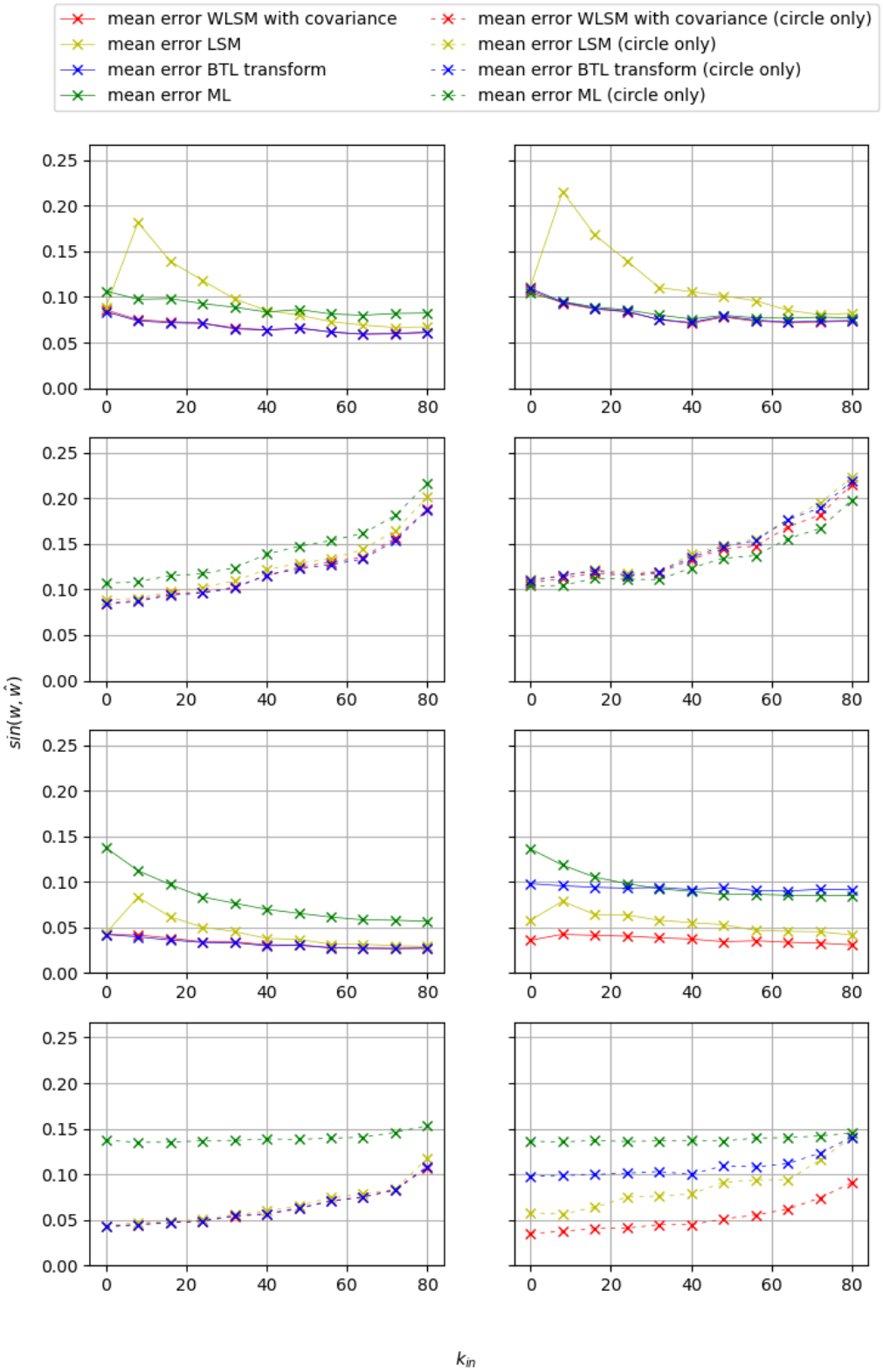


Figure E.6: Performances of the different methods applied on the different models (Rao-Kupper ($\theta = 2$), Davidson ($\theta = 1$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right with two stacked figures representing one model) for the experiment described by Figure 7.6. The parameter k_{in} denotes the number of comparisons made between non-neighbor nodes on the inside of the graph structure. The progressive increase of k_{in} hides the decrease of k_{out} between neighbor nodes (on the outside circle graph).

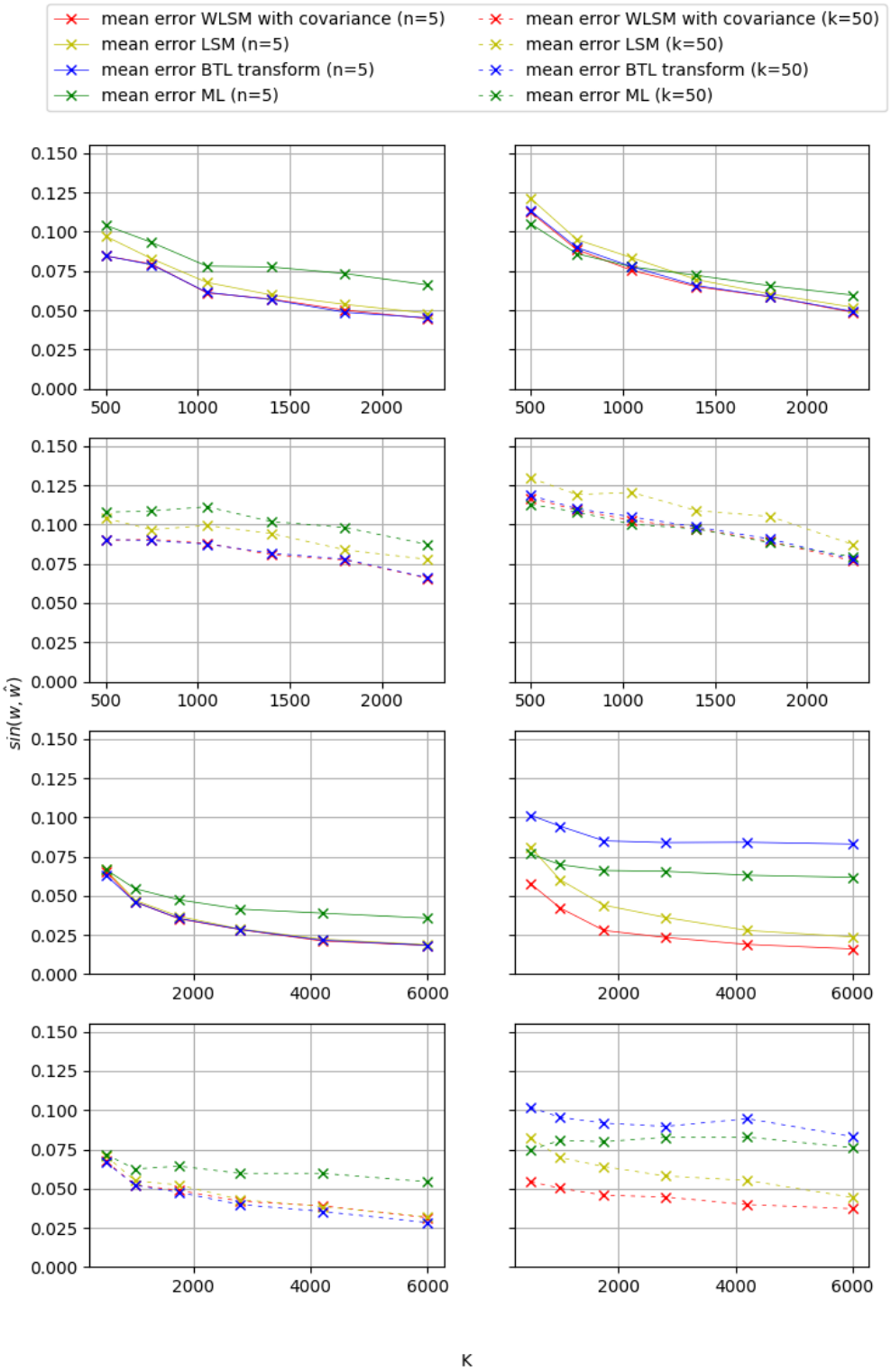


Figure E.7: Performances of the different methods applied on the different models (Rao-Kupper ($\theta = 2$), Davidson ($\theta = 1$), Plackett-Luce and Mallows-Bradley-Terry models respectively from top to bottom, left to right with two stacked figures representing one model) for the experiment described by Figure 7.8. The parameter K denotes the amount of information contained in the graph namely the number of ranking observations per pair of items multiplied by the number of pairs in the graph. For a model, the top figure shows the performance on the graph of 5 nodes with an increasing number of observations between the items while the bottom graph shows the results for a constant number of observations between item ($k = 50$) for an increasing number of items.

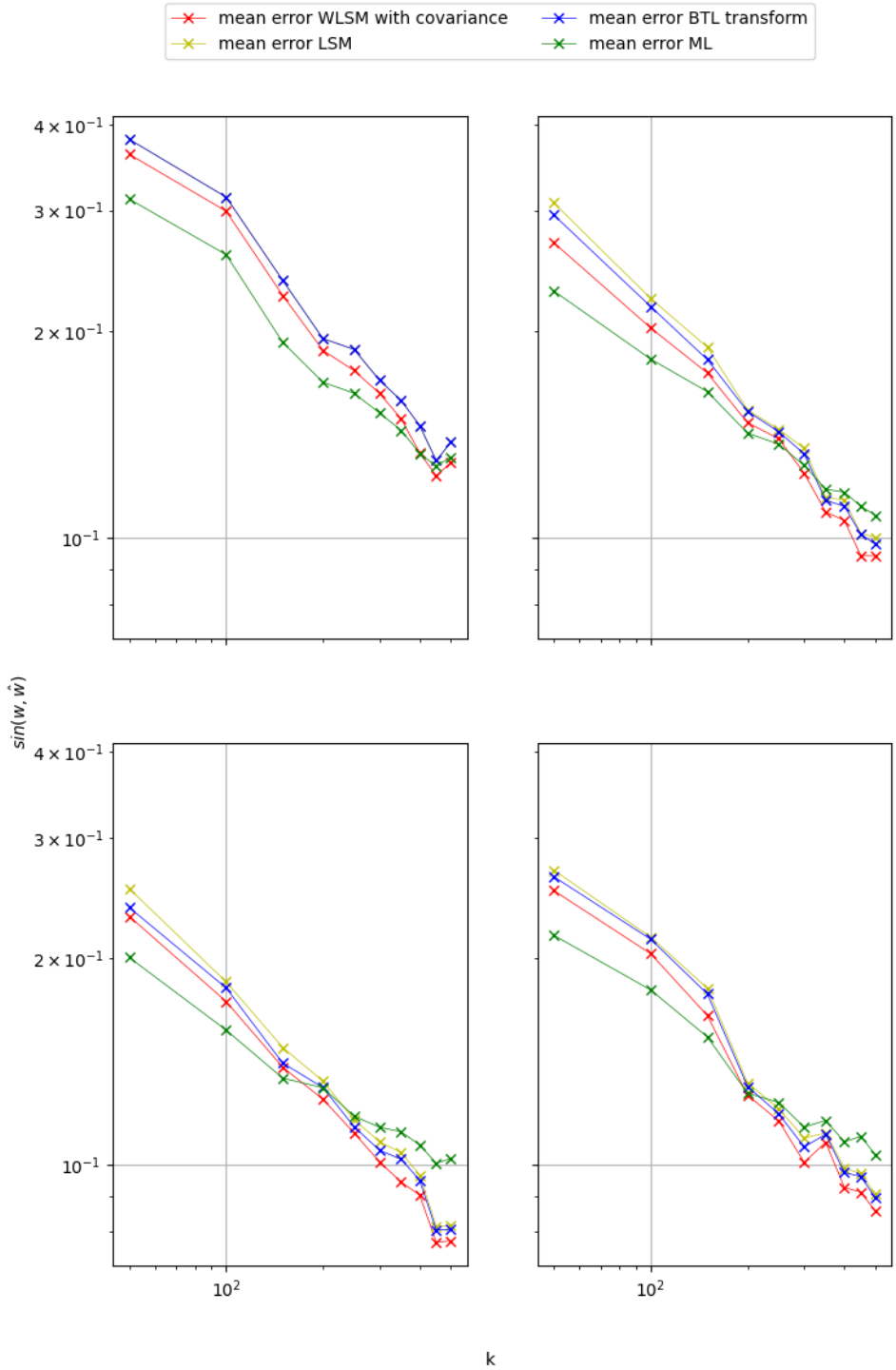


Figure E.8: Performances of the Davidson model ($\theta = 1$) for different graph structures. The upper left graph shows the results on the line (LSM performances are the same as the BTL-transform performances), the results on a circle are on the upper right figure, the lower left graph shows the performances on a 2D grid structure and the results for an Erdös-Rényi graph structure the results are shown on the lower right figure.

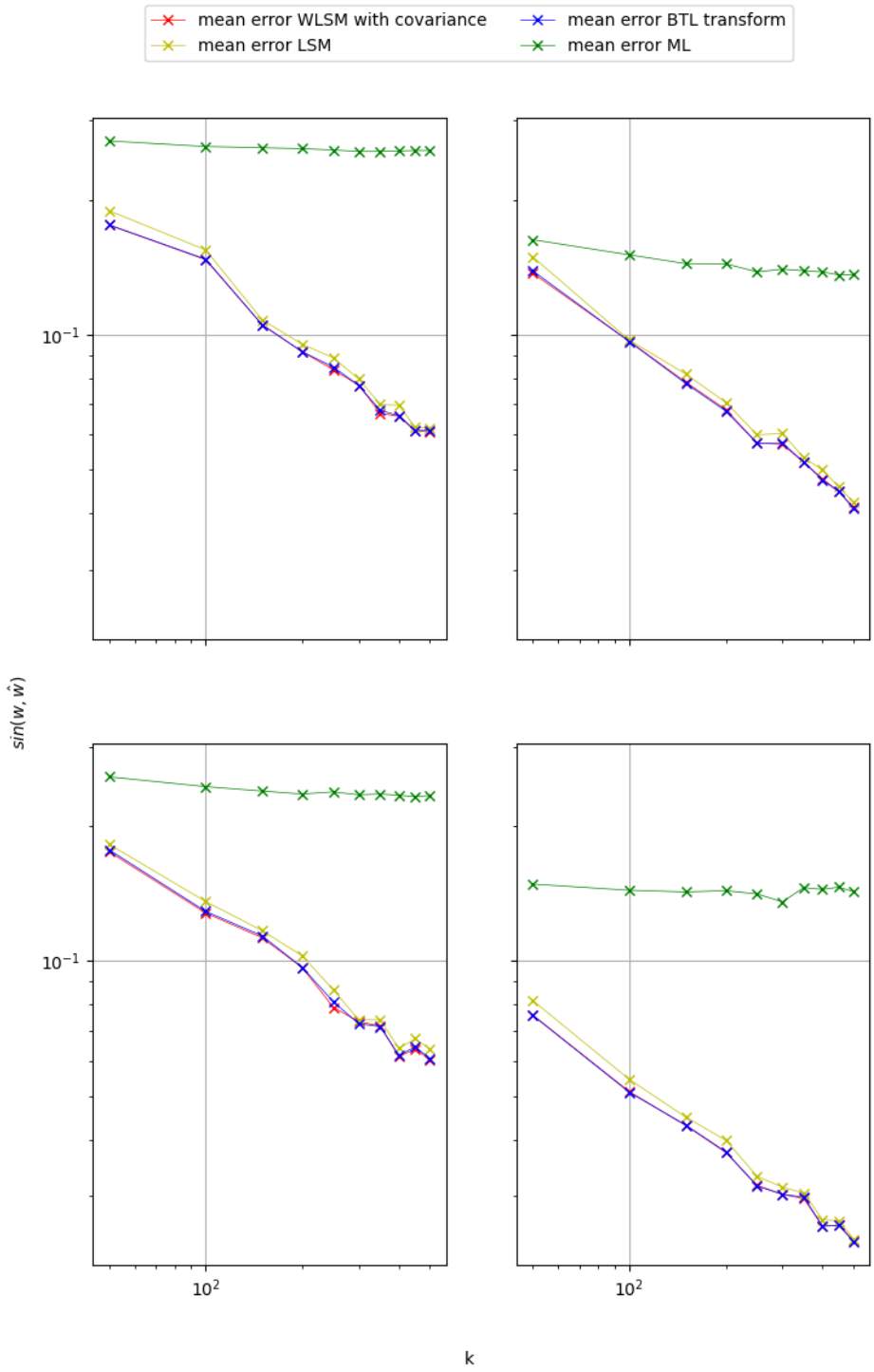


Figure E.9: Performances of the Plackett-Luce model for different graph structures. The upper left graph shows the results on the line, the results on a circle are on the upper right figure, the lower left graph shows the performances on a 2D grid structure and the results for an Erdös-Rényi graph structure the results are shown on the lower right figure.

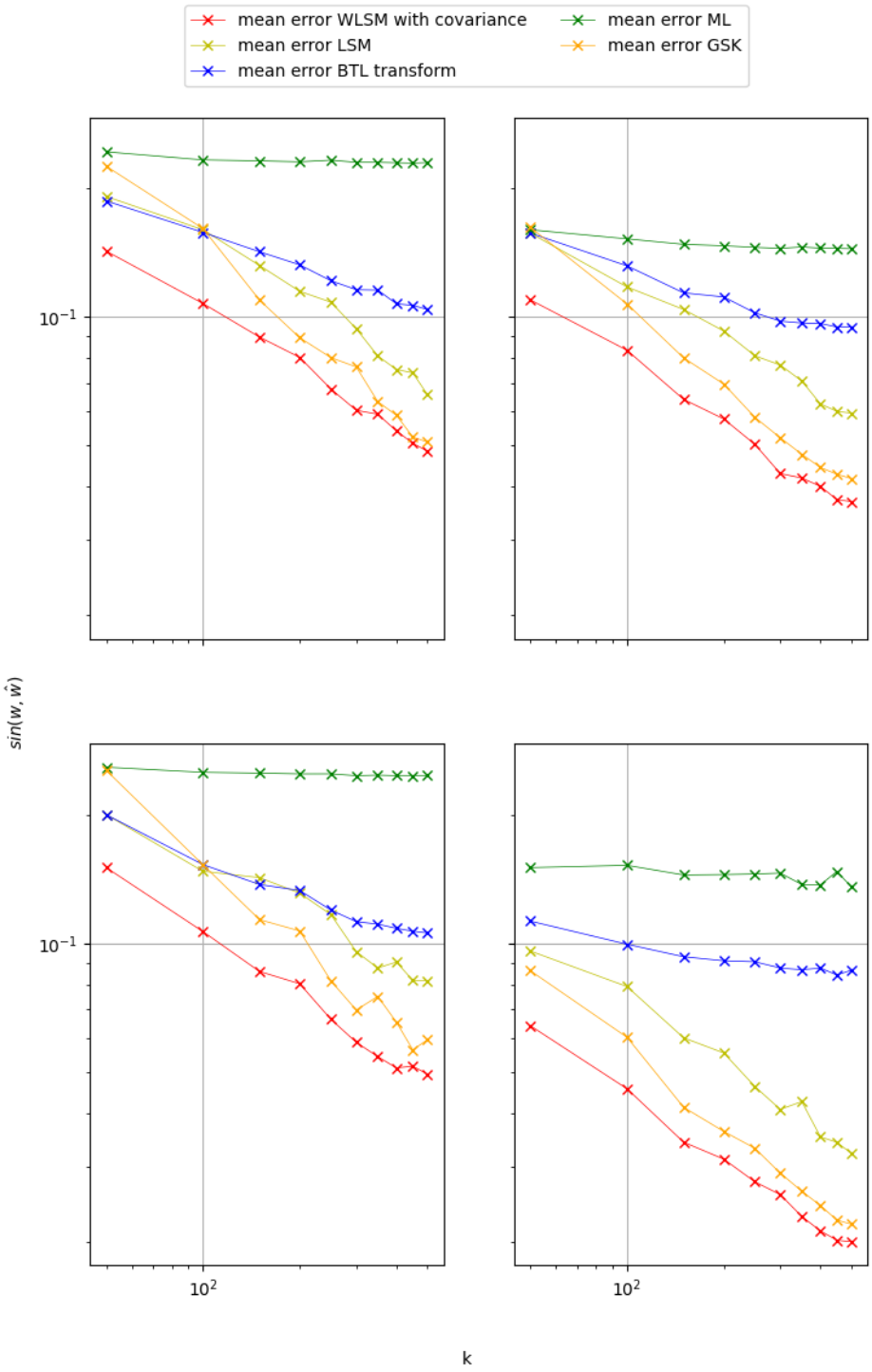


Figure E.10: Performances of the Mallows-Bradley-Terry model for different graph structures. The upper left graph shows the results on the line, the results on a circle are on the upper right figure, the lower left graph shows the performances on a 2D grid structure and the results for an Erdös-Rényi graph structure the results are shown on the lower right figure.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain
Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/epl**