SAPIENZA
UNIVERSITÀ DI ROMA

Extending parametric models for ranked data

Dipartimento di Scienze Statistiche

Dottorato di Ricerca in Statistica Metodologica – XXV Ciclo

Candidate

Cristina Mollica
ID number 955517

Thesis Advisor

Prof. Luca Tardella

A thesis submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Methodological Statistics

September 2014

Thesis defended on 17 September 2014
in front of a Board of Examiners composed by:

Prof. Alessandro De Gregorio

Prof. Chiara Gigliarano

Prof. Antonello Maruotti

Extending parametric models for ranked data
Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Website: http://www.dss.uniroma1.it/it/dipartimento/mollica-cristina

Author's email: cristina.mollica@uniroma1.it

To the memory of my mother

Ringraziamenti

Desidero, in primo luogo, esprime la mia gratitudine nei confronti del Prof. Luca Tardella, supervisore di questa tesi. È stato un privilegio ricevere la sua guida, collaborare e condividere con lui idee e stimoli durante il percorso di ricerca. Lo ringrazio soprattutto per avermi lasciato la giusta autonomia, necessaria affinché potessi fare ricerca in modo libero e consapevole, e per l'inesauribile ottimismo, che mi ha portato incredibilmente ad accettare anche gli aspetti più incerti del nostro lavoro. Grazie, Prof!

Ringrazio anche a chi, mosso da curiosità, interesse o esplicita richiesta da parte mia, ha speso parte del suo prezioso tempo a leggere parte di questa tesi e ha fornito suggerimenti utili.

Rimanendo entro le mura del dipartimento, un pensiero speciale va a tutti gli amici/colleghi che ho avuto la fortuna di conoscere durante l'avventura del dottorato. Considero un piccolo miracolo lo spirito che ha caratterizzato la nostra "convivenza forzata", improntato al confronto, alla condivisione e, immancabile, al supporto reciproco. Indimenticabili rimarranno i nostri pranzi comunitari e la voglia di ritrovarsi anche al di fuori del contesto accademico.

Sono profondamente grata, inoltre, alla mia famiglia per l'affetto e il manifestato desiderio affinché continuassi a studiare "nonostante tutto". Grazie per aver creduto sempre nelle mie capacità (più di me, sicuramente!) ed avermi incoraggiato fattivamente, quando serviva, ad andare avanti nella ricerca.

Questa tesi è interamente dedicata a mia madre. A lei va, da sempre, il mio grazie più grande.

Contents

Introduction

Choice behavior is a theme of great interest in several research areas, such as social and psychological sciences, but its investigation usually involves variables which cannot be directly observed and measured in an objective and precise manner. For this reason the evidence in choice experiments is often collected in ordinal form, that is, in terms of *ranking data*. More specifically, ranked data arise in those studies where a sample of N people is presented a finite set of K alternatives, called *items*, and is asked to rank them according to a certain criterion, typically personal preferences or attitudes. Thus, a generic ranking is the result of a comparative judgment on the competing alternatives expressed in the form of order relation. Interest in ranked data analysis is motivated, for example, by marketing and political surveys, where items could be consumer goods, political candidates or goals, but also by psychological and behavioral studies consisting, for instance, in the ordering of words/topics according to the perceived association with a reference subject. However, ranking data are not limited to the inquiry of preferences or attitudes of a target population. Another typical context that naturally gives rise to rankings is sports, such as national soccer championships, as well as horse or car races, where players or teams compete and the final outcome is a ranking among competitors.

Ranked data analysis has been addressed from numerous perspectives, as revealed by a wide and consolidated literature on this topic. This thesis focuses on the probabilistic modeling of ranking data and, after reviewing the main contributions on the construction of statistical models for random rankings, develops some original extensions of a popular parametric distribution.

The thesis is organized as follows. Chapter 1 formalizes notation and provides a structured overview of several methodological strategies adopted in building parametric modeling for rankings. The review describes the basic approaches developed in the literature to conceive non-uniform models, which can be classified in four main categories: (i) order statistics models, (ii) models based on paired comparisons, (iii) distance-based models and (iv) stagewise models. We will then concentrate on a parametric distribution belonging to the last class, known as *Plackett-Luce model* (PL). Probability distributions based on a sequential construction of the ranking, such as the PL, implicitly suppose that preferences are expressed with the canonical *forward procedure*, meaning that the judge proceeds from the elicitation of her best choice up to the worst one. In spite of various attempts to improve the description of ranked data, yet the rank assignment order has not, to our knowledge, received explicit consideration in any model setup process, although any other order for the rank assignment process is admissible and potentially leads to different results. This aspect has inspired us to extend the PL relaxing the conventional forward assump-

tion and representing the order of the rank assignment scheme with an additional free parameter in the model, referred to as *reference order*. The novel generalization, named *Extended Plackett-Luce model* (EPL), is introduced in the last part of Chapter 1. A characterizing property of the ordinary PL is employed to formally prove the actual greater flexibility of the proposal.

Chapter 2 is completely dedicated to the estimation procedures for ranking models within the maximum likelihood inferential framework. We describe how to solve the inferential issue of maximizing the likelihood function for the novel EPL over a mixed-type parameter space, due to the discreteness of the reference order, thanks to the adaptation and combination of different estimation devices for ranking models. For the continuous parameters we apply the *Minorization/Maximization* (MM) algorithm, which is an iterative optimization technique based on the replacement of the original objective function with a more tractable minorizing surrogate function. For the discrete parameter, instead, we implement a local search, constraining the optimization step within a fixed distance from the current estimate of the reference order. To evaluate the sensitivity of the algorithm w.r.t. the choice of a particular distance in the local search step, we focus on two frequently used metrics for rankings and compare the corresponding estimation performances. Moreover, in order to address the common situation of unobserved sample heterogeneity and increase the applicability of the EPL, we also consider the natural extension of the novel model in the mixture model setting. In this framework the likelihood maximization requires the derivation of a hybrid procedure, called EMM algorithm, which integrates the standard Expectation Maximization algorithm with the aforementioned MM procedure.

In Chapter 3 we verify the practical utility of the EPL with an application to the real LFPD data set. This data set comes from a bioassay experiment and collects the binding measurements of human blood exposed to 11 partially overlapping fragments of the HER2 oncoprotein. Raw quantitative outcomes have been obtained from three different disease groups but, for reasons due to the numerical instability of measurements and to the absence of universally accepted methods of rescaling the original data, we are interested in verifying the possible usefulness of the underlying ordinal information as a more robust and unambiguously defined evidence of sample heterogeneity. Specifically, we address the heterogeneous nature of the experimental units via model-based clustering and compare the performance of the mixture model using the new distribution as mixture components with alternative mixture models for random rankings.

The successful application of the EPL to bioassay data encouraged us to explore further PL generalizations in different directions. These ideas include the possibility to combine the novel EPL with the popular *Benter model* (BM). The BM extends the PL with the introduction of additional parameters that account for variable selection accuracy over the stages of the ranking process. As described in Chapter 4, the EPL and the BM move from substantially different but complementary attributes of the ranking procedure and their merging can add further flexibility to the PL. We name this second proposal *Benterized Extended Plackett-Luce model* and detail how to perform maximum likelihood estimation for both the homogeneous population case and the finite mixture framework.

As final contribution, Chapter 5 illustrates the extension to the finite mixture

context of a Bayesian device for PL inference recently introduced in the literature. We describe an efficient way to incorporate the latent group structure in the data augmentation approach and how to interpret previous maximum likelihood procedures as special instances of the proposed Bayesian analysis. We then test the computational effectiveness and efficiency of both the maximum *a posteriori* estimation and the GS algorithm under multiple simulation scenarios focusing, in particular, on the identifiability problems that can affect the results of the MCMC technique. In this regard, we discuss the application of several relabeling algorithms to the MCMC samples, aimed at solving the label-switching issue, and examine the resulting posterior inference. We finally illustrate the Bayesian inference for the PL mixture with an application to a real data set concerning taste preferences of different hamburger cooking styles.

The thesis ends with concluding remarks and proposals for future developments involving, in particular, the idea to implement also the EPL mixture within the Bayesian paradigm.

Chapter 1

Statistical models for ranked data

1.1 Notation and basic definitions

Before reviewing the main approaches for the probabilistic modeling of ranked data, it is convenient to fix some notation. Formally, a *full* (or *complete*) *ranking* π is a bijective mapping of a finite set $I = \{1, \ldots, K\}$ of labeled *items* into a set of *ranks* $R = \{1, \ldots, K\}$, that is

$$\pi : I \to R,$$

resulting from the attribution of a rank to each item. The whole bijective mapping can be represented in terms of an ordered K-tuple $\pi = (\pi(1), \ldots, \pi(K))$, where positions of the components refer to items and entries give the corresponding assigned ranks. In other words, $\pi(i)$ must be read as the rank attributed to the i-th item. The underlying convention is that if $\pi(i) < \pi(i')$, then item i is ranked higher than item i', and hence preferred to it.

In the literature, it is common to distinguish between a *full* and a *partial* (or *incomplete*) *ranking* where, in the latter case, the rank assignment process is not completely carried out. This happens, for instance, when a judge expresses only her first t preferences out of K items ($t < K$), producing the so-called top-t partial ranking. Top-t rankings are just one type in the wide variety of incomplete data that can emerge from a ranking experiment (see Lebanon & Mao (2008)) and the situation where both full and partial rankings are involved in the statistical analysis is usually referred to as *heterogeneous ranked data analysis*. In the past it was a usual practice to discard incomplete rankings from the analysis, with the consequent sample size reduction and loss of inferential accuracy. In the last decades several proposals to extend the standard ranking models have appeared in the literature and a substantial evidence gain resulting from the treatment of heterogeneous ranked data has been emphasized. In the first part of the thesis we suppose that the bijection π is completely observed, so that partial rankings are not permitted because of surjectivity of the mapping and ties are not allowed due to injectivity. In Chapter 5, where a novel Bayesian modeling proposal is presented, we relax this assumption to account also for top-t partial observations in the sample.

The inverse $\pi^{-1} = (\pi^{-1}(1), \ldots, \pi^{-1}(K))$ of a ranking π is called *ordering*. Positions of the components in π^{-1} refer to ranks and elements correspond to items. Hence, $\pi^{-1}(j)$ is the item ranked in the j-th position. In order to avoid confusion with π, we will henceforth make explicit use of the inverse function notation to denote the corresponding ordering $\pi^{-1} : R \to I$. We remind also that a ranking admits a matricial representation Π with binary elements defined as follows

$$\Pi_{ij} = \begin{cases} 1 & \text{if } \pi(i) = j, \\ 0 & \text{otherwise} \end{cases} \qquad i, j = 1, \ldots, K.$$

The binary matrix Π is doubly stochastic and its inverse $\Pi^{-1} = \Pi^T$ is also a binary doubly stochastic matrix returning the matricial format for the corresponding ordering.

We denote the set of all $K!$ possible rankings with $\mathcal{S}_K$, which is identified with the symmetric group of permutations endowed with the composition operation $\circ$. The composition $\pi\sigma^{-1} = \pi \circ \sigma^{-1} = (\pi(\sigma^{-1}(1)), \ldots, \pi(\sigma^{-1}(K)))$, for instance, indicates ranks under π relative to the items ranked $1, \ldots, K$ by σ. In the symmetric group this is formally defined as the mapping σ^{-1} which acts on the right of π, to stress that the composition is not in general commutative. Reversing the order in the composition, in fact, gives $\sigma^{-1}\pi = \sigma^{-1} \circ \pi = (\sigma^{-1}(\pi(1)), \ldots, \sigma^{-1}(\pi(K)))$, which lists items receiving from σ those ranks that π has attributed to items $1, \ldots, K$. Geometrically speaking the elements of $\mathcal{S}_K$ can be represented as the vertices of a *polytope*, i.e., of the convex hull of the points $\pi \in \mathcal{S}_K \subset \mathbb{R}^K$. Thompson (1993) suggested this representation to visualize the distribution of a sample $\underline{\pi} = \{\pi_1, \ldots, \pi_s, \ldots, \pi_N\}$ of N rankings, placing a ball at each vertex of the polytope with diameter proportional to the corresponding frequency observed in the sample (Figure 1.1). In addition to a graphical inspection, feasible only for a small number K of items, it could be useful to perform a preliminary exploratory analysis on the ranking data set or on specific parts thereof. As examples of descriptive statistics one can mention *the first-order marginals matrix* $\hat{M}$, whose generic element $\hat{M}_{ij}$ is defined as

$$\hat{M}_{ij} = \frac{1}{N} \sum_{s=1}^{N} I_{[\pi_s(i)=j]} \qquad i, j = 1, \ldots, K,$$

where $I_{[E]}$ is the indicator function of the event E such that $I_{[E]} = 1$ if the event E occurs and $I_{[E]} = 0$ otherwise. Thus, $\hat{M}_{ij}$ indicates the observed relative frequency that item i is ranked j-th. Comparing pairs of items rather than considering the marginal distribution for each column of the data set, one can construct the *pairs matrix* $\hat{P}$ where

$$\hat{P}_{ii'} = \frac{1}{N} \sum_{s=1}^{N} I_{[\pi_s(i)<\pi_s(i')]} \qquad 1 \leq i < i' \leq K$$

is the proportion in the sample preferring item i to item i'. Obviously, in the absence of ties one has $\hat{P}_{i'i} = 1 - \hat{P}_{ii'}$. Another summary statistics is the *mean* or *average rank vector* given by

$$\overline{\pi} = \frac{1}{N} \sum_{s=1}^{N} \pi_s,$$

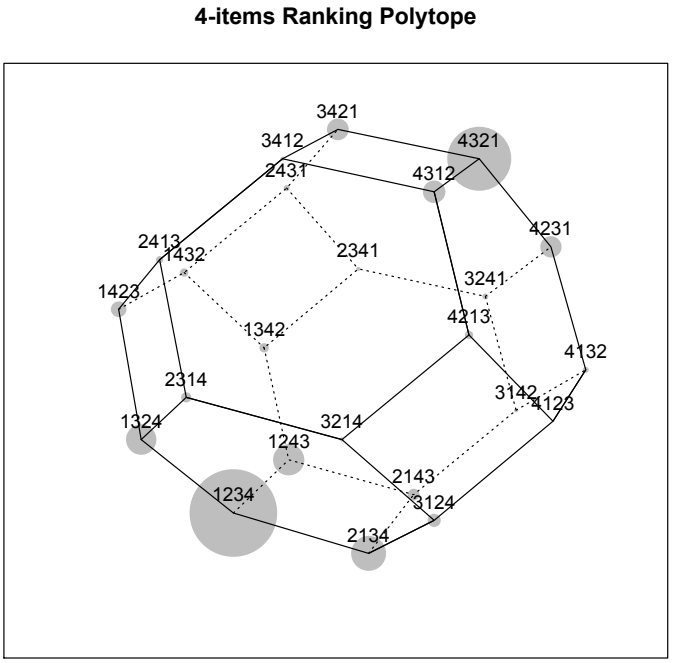

Figure 1.1. 3-D representation of the ranking polytope for $K = 4$ items. Rankings are represented as the vertices of the polytope and segments connect adjacent ordered sequences in terms of the Kendall distance. Ball radius is proportional to the sample frequency of each ranking.

where the generic element $\overline{\pi}(i)$ expresses the mean rank of item i. Note that $\overline{\pi}$ is not properly a ranking. The average rank vector $\overline{\pi}$ can be easily obtained from the $K \times K$ matrix $\hat{M}$ using the following relation

$$\overline{\pi} = \hat{M}e,$$

where $e = (1, \ldots, K)$ is the *identity ranking* in $\mathcal{S}_K$. The above summaries are implemented by the function `destat` of the R package `pmr`, whereas the command instruction `rankplot` displays the empirical distribution on the ranking polytope in the presence of $K = 3$ or $K = 4$ items.

1.2 Probability models for random rankings

In this section we give a brief review of ranked data modeling. It is not intended to be an exhaustive outline but offers some useful background of the most frequently used distributions for rankings. For a more systematic review see Marden (1995) and Critchlow et al. (1991).

The most general statistical model, referred to as *saturated model* (SM), is the collection of all discrete distributions for random rankings, identified with the whole $(K! - 1)$-dimensional simplex $\mathcal{P}(\mathcal{S}_K)$ and parameterized by the $K! - 1$ probabilities of each ordered sequence. This very general class includes elements which play a special role in ranking modeling: the *uniform* (or *null*) *model* (UM), represented by the single flat distribution which assigns equal probability to each ranking, and the

degenerate models (DM), which conversely concentrate all the probability mass on a single ranking. Geometrically, the UM is the central point of $\mathcal{P}(\mathcal{S}_K)$ and the DM coincides with its vertices. The SM incorporates every possible statistical model, whereas the UM and the DM are usually obtained as specific members of almost all common statistical models for random ranking. In this sense the UM and the DM should be regarded as reference probability distributions.

Although the SM allows for the maximum degree of flexibility, the fast-growing dimension of the ranking space makes it intractable and cumbersome to interpret even with a relatively small number K of items. An empirical example of such a difficulty may be the inference of the distribution parameters from the observed orderings of web-pages produced by different search engines, for which K is usually very large. These practical limitations have motivated the introduction of simplifying assumptions on the ranking process and justify the wide assortment of restricted parametric models developed in rank data theory. In this regard we can recover also in the ranking literature the more general issue of statistical modeling, whose objective is the definition of interpretable and useful parametric representations, able to capture and efficiently synthesize the relevant aspects of empirical phenomena. Following the review in Critchlow et al. (1991), the basic approaches developed in the literature to define non-uniform models can be classified in four main categories: (i) order statistics models, (ii) models based on paired comparisons, (iii) distance-based models and (iv) stagewise models. These alternative strategies do not have to be interpreted as unrelated compartments in modeling the ranking process. Some effort in our review, in fact, will be devoted to clarify the meaning of the underlying assumptions within specific approaches but also the overlaps among different parametric families.

1.2.1 Ordered statistics models

One of the main streams in the development of ranked data modeling is the approach based on the order statistics. Thurstone (1927) proposes to use an auxiliary continuous multivariate model to derive a model for random rankings as a by-product of the distribution of the order statistics. The *Thurstone model* (TM) assumes the existence of a unobserved quantitative mechanism underlying the ranking process, such that each item i is associated with a continuous latent random variable (r.v.) W_i, also named *score*. The score should be intended as a latent item feature, measurable on a unidimensional scale and on which the comparative judgment is based. Examples of unidimensional scores are the unrecorded arrival times of drivers in a race or any possible preference/liking measure towards items. In this perspective the derivation of a model on $\mathcal{S}_K$ induced by a hypothesized parametric joint distribution for the W's is straightforward if one considers the order of the item scores as follows

$$\mathbf{P}(\pi) = \mathbf{P}(W_{\pi^{-1}(1)} < \cdots < W_{\pi^{-1}(K)}) \qquad \pi \in \mathcal{S}_K. \tag{1.2.1}$$

It suffices to set $\pi(i) = \text{rank}(W_i)$ in $\{W_1, \ldots, W_K\}$ for all $i = 1, \ldots, K$ to switch from continuous to ordinal information. Postulating that the W_i's are independent and normally distributed with different means but same variances, the expression in (1.2.1) translates into the *Thurstone-Mosteller-Daniels model* (see Mosteller

(1951) and Daniels (1950)). More general versions of this approach relax the hypothesis of homoscedasticity, of independence, or contemplate other parametric laws for the item scores. Some contributions in this direction can be found in Luce (1959) and Henery (1983), who respectively considered the Gumbel and the Gamma distributions with equal shape parameters to model the latent W's. Following the terminology in Train (2003), Thurstone models are also known as *random utility models*.

1.2.2 Ranking models based on paired comparisons

The TM has an interesting connection with another approach to construct a random ranking based on the so-called pairwise comparisons. Originally, in fact, the TM involved only two items and Daniels (1950) contributed to extend it for the ordering of $K > 2$ alternatives. For a detailed review on the literature of paired comparisons the reader is referred to Bradley & Terry (1952), Bradley (1976), Bradley (1984) and David (1988). Suppose the items in I are compared and ordered in a pairwise manner, for a total of $K(K-1)/2$ comparisons. Let us define with the binary r.v. $X_{ii'} \sim \text{Bern}(\rho_{ii'})$ the preference between item i and i' in the paired comparison, so that

$$X_{ii'} = \begin{cases} 1 & \text{if item } i \text{ is preferred to item } i', \\ 0 & \text{otherwise} \end{cases} \qquad 1 \leq i < i' \leq K.$$

A generic set of paired comparisons, however, is not necessary consistent with the definition of a ranking. In fact, circularities of the type $\pi(1) < \pi(2) < \pi(3) < \pi(1)$ are allowed in pairwise comparison modeling, whereas are not permitted in a ranking elicitation. It follows that a conditioning argument is needed in order to construct a model for rankings after converting them into the suitable set of pairwise preferences. Assuming that all the $K(K-1)/2$ comparisons are drawn independently and governed by the probabilities $\rho_{ii'}$, for the final result of the random ranking to be valid it is necessary that the sequence of paired preferences does not contain any circularity, otherwise it is discarded and the comparisons are repeated until no circularity is present. In this setting the probability of each ranking turns out to be

$$\mathbf{P}(\pi|\underline{\rho}) = \frac{\prod_{i<i'} \rho_{ii'}^{x_{ii'}(\pi)}(1-\rho_{ii'})^{1-x_{ii'}(\pi)}}{\sum_{\pi \in \mathcal{S}_K} \prod_{i<i'} \rho_{ii'}^{x_{ii'}(\pi)}(1-\rho_{ii'})^{1-x_{ii'}(\pi)}} \qquad \pi \in \mathcal{S}_K,$$

which is known in the literature as the *Babington Smith model* (BSM). It was originally proposed by Babington Smith (1950) and is indexed by the $K(K-1)/2$ parameters collected in $\underline{\rho} = (\rho_{ii'})_{i<i}$. We remind that, since ties are not permitted, $\rho_{i'i} = 1 - \rho_{ii'}$.

Setting special forms for the ρ's, popular subclasses of the BSM can be derived. Bradley & Terry (1952) introduced item parameters $p_i > 0$ for all $i = 1, \ldots, K$ reflecting the skill rate of each item and constrained the paired comparison probabilities as follows

$$\rho_{ii'} = \frac{p_i}{p_i + p_{i'}}. \tag{1.2.2}$$

Equality (1.2.2) is the basic equation of the well-known *Bradley-Terry model* (BT), which the authors applied only to paired comparison data. Mallows (1957) suggested to substitute expression (1.2.2) in the BSM, leading to the *Mallows-Bradley-Terry model* for rankings. The same author proposed other simplifications which reduce the BSM to specific distance-based ranking models, extensively described later in Section 1.2.4. We conclude this section stressing that, in the presence of only two items i and i', the TM corresponds to the BT when the scores W_i and $W_{i'}$ follow the Gumbel distribution. The same parametric setup for $K > 2$ leads to the Plackett-Luce ranking model discussed in the next section.

1.2.3 The Plackett-Luce model and related extensions

When a judge proceeds to elicit from her best choice (rank 1) up to the worst one (rank K), we say that she is eliciting according to the so-called *forward ranking process*; the inverse ranking procedure is named *backward ranking process*. This formal definition has been originally introduced in Fligner & Verducci (1988) but, to our knowledge, the rank assignment scheme has not received an explicit consideration in a model setup in the attempt to improve the description of random ranked data. Obviously, any other order for the rank assignment process is admissible and potentially leads to different models. As detailed in Section 1.3, this aspect has inspired us to expand an existing and well-known parametric ranking model, that is the *Plackett-Luce model* (PL). It is a very popular parametric family of ranking distributions and can be considered in turn as an extension of the BT to the context of multiple (more than two) repeated item comparisons. Its name arises from both independent contributions supplied by Luce (1959) and Plackett (1975). The monograph in Luce (1959) provides an in-depth theoretical description of the individual choice behavior based on a general axiom system, whereas Plackett (1975) derived the PL in the context of horse races. Its probabilistic representation moves from the decomposition of the ranking process in a finite sequence of independent stages, one for each rank that has to be assigned, combined with the underlying assumption of standard forward procedure on the ranking elicitation. In fact, a ranking can be elicited through a series of sequential comparisons in which a single item is preferred to all the remaining alternatives and, after being selected, is removed from the next comparisons. For this reason, the PL is said to belong to the family of *stagewise ranking models*. Specifically, the PL probability distribution is completely specified by the so-called *support parameter* vector $\underline{p} = (p_1, \ldots, p_K)$, where $p_i > 0$ for all $i = 1, \ldots, K$ and $\sum_{i=1}^{K} p_i = 1$. Note that in the present PL formulation the support parameters are constrained to add up to one. This restriction is introduced to avoid unidentifiability due to possible multiplication by an arbitrary positive constant. The generic parameter component p_i expresses the probability that item i is selected at the first stage of the ranking process and hence preferred among all other items. The probability of choosing item i at lower preference levels $t > 1$ is proportional to its support value p_i. Since the set of available items in the sequence of random selections is reduced by one element after each step, the computation of the choice probabilities at each stage requires suitable normalization of the support probabilities w.r.t. the set of remaining items. It follows that under the PL the

probability of the random ordering π^{-1} is

$$\mathbf{P}(\pi^{-1}|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(t)}}{\sum_{\nu=t}^{K} p_{\pi^{-1}(\nu)}} \qquad \pi^{-1} \in \mathcal{S}_K. \tag{1.2.3}$$

The special case $p_i = 1/K$ for all $i = 1, \ldots, K$ corresponds to the UM over $\mathcal{S}_K$.

The vase model metaphor, originally introduced by Silverberg (1980), is an alternative way to interpret the above random stagewise item selections and a useful representation in order to understand the PL extensions already developed in the literature (see Marden (1995) for a review). Let us consider a vase containing item-labeled balls such that the vector $\underline{p}$ expresses the starting vase composition. The vase differs from an urn simply because the former contains an infinite collection of balls in order to allow arbitrary proportions in the simplex. At the first stage a ball is drawn and the corresponding item is ranked first. At the second stage another ball is drawn from the vase: if its label is different from $\pi^{-1}(1)$ the rank 2 is then assigned to the corresponding item, otherwise the ball is put back in the vase and the drawing is repeated until a distinct item is chosen and then ranked in the second position. The multistage experiment ends when there is only one item not yet selected and this is automatically ranked last. Put in other words, PL can be regarded as an urn sampling scheme without replacement where the vector $\underline{p}$ describes the inclusion probabilities of each item-labeled ball. The probability of a generic sequence of drawings turns out to be (1.2.3). In such a scheme the vase configurations (inclusion probabilities) are constant over all stages and stagewise interactions among items are not contemplated. A first attempt to generalize this basic scheme consists in retaining the absence of item interactions and allowing a varying vase composition among stages, as formalized in Silverberg (1980). In this model setting the support parameters become stage-dependent, that is p_{ti} for $t = 1, \ldots, (K-1)$ and $i = 1, \ldots, K$. Setting the special form $p_{ti} = p_i^{\alpha_t}$ one obtains the *Benter model* (BM) introduced by Benter (1994), that is

$$\mathbf{P}(\pi^{-1}|\underline{p}, \underline{\alpha}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(t)}^{\alpha_t}}{\sum_{\nu=t}^{K} p_{\pi^{-1}(\nu)}^{\alpha_t}} \qquad \pi^{-1} \in \mathcal{S}_K. \tag{1.2.4}$$

The BM is then characterized by the additional vector $\underline{\alpha} = (\alpha_1, \ldots, \alpha_K)$ of the *dampening parameters* with $0 \leq \alpha_t \leq 1$ for all $t = 1, \ldots, K$. As apparent from the probabilistic definition (1.2.4), the element of $\underline{\alpha}$ perturb the stagewise stochastic selections given by the sequential normalizations of the item supports. These parameters accommodate for the possible different degree of accuracy with which the choice at each selection stage is made. The choice accuracy, for example, may depend on the cardinality of the item set I: if K is large, the ranker may have clear preferences about her best choices and order, instead, the remaining alternatives with less accuracy. The PL corresponds to the BM with $\alpha_t = \alpha = 1$ for all $t = 1, \ldots, K$, meaning that all ranks are attributed with the maximum degree of care. In order to overcome overparametrization problems, one assumes $\alpha_1 = 1$ and $\alpha_K = 0$ in the inferential analysis (see Gormley & Murphy (2008a) and Gormley & Murphy (2008b)). Moreover, one can also relax the non-interaction hypothesis, so that the vase configuration at each stage relies on the previous selected items.

Indeed Plackett (1975) defined a hierarchy of further PL extensions. Marden (1995) refers to such generalizations as *Lag L models*, where $L = 0, \dots, K-2$ indicates that the vase at stage t depends on the previous choices only through the last L selected items $\{\pi^{-1}(t-L), \dots, \pi^{-1}(t-1)\}$. The Lag 0 model, also called in Plackett (1975) first-order model, coincides with the ordinary PL; the Lag 1 model is such that at each choice step t the vase composition depends only on the item $\pi^{-1}(t-1)$. In general, the total number of parameters in the Lag L model is given by

$$\begin{aligned}\sum_{l=0}^{L} \frac{K!}{(K-l)!}(K-l-1) &= \sum_{l=0}^{L}\left(\frac{K!}{(K-l-1)!} - \frac{K!}{(K-l)!}\right) = \frac{K!}{(K-L-1)!} - 1 \\ &= K(K-1)\cdots(K-L) - 1,\end{aligned}$$

thus the $L = K-2$ model covers the whole SM.

1.2.4 Distance-based models

Another fundamental class of parametric distributions is the so-called *distance-based model* (DB). Roughly speaking, the DB can be interpreted as the analogue of the normal distribution on the finite discrete space $\mathcal{S}_K$. In fact, it is an exponential location-scale model indexed by a discrete location parameter $\sigma \in \mathcal{S}_K$, called *modal* or *central ranking*, and a non-negative real *concentration parameter* $\lambda \in \mathbb{R}_0^+$. Each distribution in a DB model has the following form

$$P(\pi|\sigma,\lambda) = \frac{1}{Z(\lambda)} e^{-\lambda d(\pi,\sigma)} \qquad \pi \in \mathcal{S}_K, \tag{1.2.5}$$

where $Z(\lambda) = \sum_{\pi\in\mathcal{S}_K} e^{-\lambda d(\pi,\sigma)}$ is the normalization constant and d is a metric on $\mathcal{S}_K$. The probability mass function in (1.2.5) attains its maximum at $\pi = \sigma$ and decreases as the distance from σ increases. Under (1.2.5) rankings at the same distance from the modal sequence σ are equally probable. The central ranking σ expresses the so-called *consensus* in the population, whereas the concentration parameter λ calibrates the effect of d on the probability of the ranking: the higher the value of λ, the more concentrated the distribution around its mode (Figure 1.2). Hence, when $\lambda \to +\infty$, equation (1.2.5) becomes the DM at $\pi = \sigma$; conversely, when $\lambda = 0$ it turns out to be the UM.

Changing the distance measure d in (1.2.5), one can define different families of parametric distributions for ranked data. Following Diaconis (1988), a function $d : \mathcal{S}_K \times \mathcal{S}_K \to \mathbb{R}_0^+$ is a distance between rankings if it satisfies the usual three properties:

- *identity* $\quad \pi,\sigma \in \mathcal{S}_K \quad d(\pi,\sigma) = 0 \iff \pi = \sigma,$
- *symmetry* $\quad \pi,\sigma \in \mathcal{S}_K \quad d(\pi,\sigma) = d(\sigma,\pi),$
- *triangle inequality* $\quad \pi,\sigma,\delta \in \mathcal{S}_K \quad d(\pi,\delta) \le d(\pi,\sigma) + d(\sigma,\delta)$

and the additional fourth condition

- *right-invariance* $\quad \pi,\sigma,\delta \in \mathcal{S}_K \quad d(\pi,\sigma) = d(\pi\delta^{-1}, \sigma\delta^{-1}).$

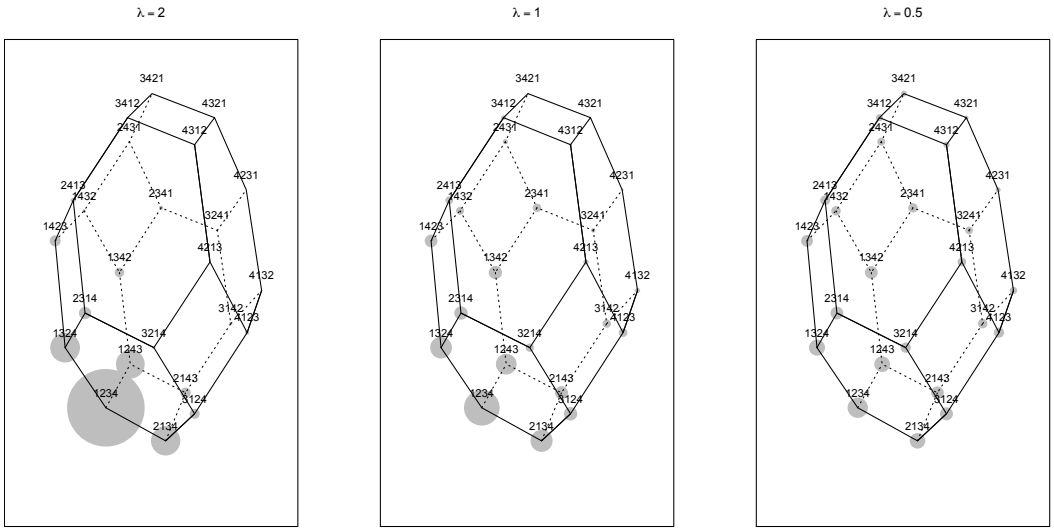

Figure 1.2. Three distance-based distributions with $d = d_K$ and modal ranking equal to $e = (1, 2, 3, 4)$ for decreasing values of the concentration parameter, respectively equal to $\lambda = 2$ (left), $\lambda = 1$ (center) and $\lambda = 0.5$ (right). The dispersion of the probability mass function on $\mathcal{S}_4$ increases as the value of λ decreases.

Right-invariance (or *label-invariance*) guarantees the desirable property of invariance of d w.r.t. arbitrary relabeling of items. Examples of metrics for rankings satisfying above properties are

- the *Kendall distance*

$$d_K(\pi, \sigma) = \sum \sum_{1 \le i < i' \le K} I_{[(\pi(i) - \pi(i'))(\sigma(i) - \sigma(i')) < 0]},$$

- the *Spearman distance*

$$d_S(\pi, \sigma) = \sqrt{\sum_{i=1}^{K} (\pi(i) - \sigma(i))^2},$$

- the *Spearman Footrule*

$$d_F(\pi, \sigma) = \sum_{i=1}^{K} |\pi(i) - \sigma(i)|,$$

- the *Cayley distance*

$$d_C(\pi, \sigma) = \text{minimum number of transpositions needed to transform } \pi^{-1} \text{ into } \sigma^{-1},$$

- the *Hamming distance*

$$d_H(\pi, \sigma) = \#\{i = 1, \ldots, K : \pi(i) \neq \sigma(i)\},$$

  i.e., the number of items ranked differently by π and σ,

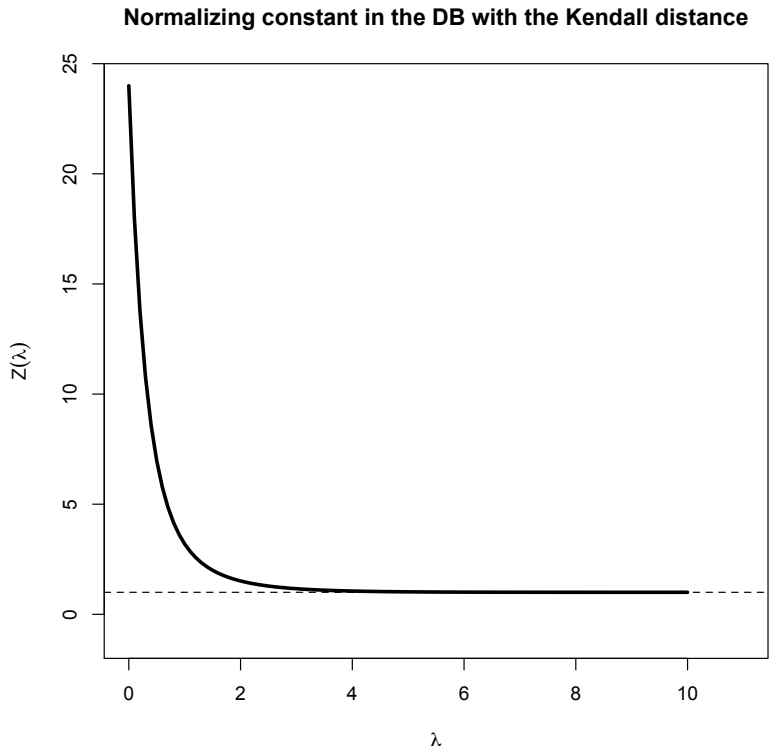

Figure 1.3. Normalizing coefficient of the distance-based model with $d = d_K$ as function of the concentration parameter λ in the case $K = 4$.

- the *Maximum distance*

$$d_M(\pi, \sigma) = \max_{1 \le i \le K} |\pi(i) - \sigma(i)|,$$

- the *Ulam distance*

$$d_U(\pi, \sigma) = K\text{-maximal number of items ranked in the same order by } \pi \text{ and } \sigma.$$

One of the most frequently used metrics is d_K, which counts the number of pairwise disagreements, i.e., the pairs of items with relative discordant order under π and σ. It is also equal to the minimum number of adjacent transpositions needed to transform π^{-1} into σ^{-1}. Relaxing the adjacency requirement one has the Cayley distance d_C, meaning that the inequality $d_C \le d_K$ always holds. Moreover, the well-known Kendall and Spearman rank correlation coefficients are obtained respectively from d_K and d_S^2 with an appropriate rescaling as follows

$$\text{corr}(\pi, \sigma) = 1 - 2\frac{d(\pi, \sigma)}{d_{\max}},$$

where $d_{\max} = \max\{d(\pi, \sigma) : \pi, \sigma \in \mathcal{S}_K\}$. For a more complete and detailed description of the metrics on rankings, the reader is referred to Critchlow (1985) and Diaconis (1988).

Due to the right-invariance of d, the normalizing constant in (1.2.5) does not depend on σ:

$$Z(\lambda) = \sum_{\pi \in \mathcal{S}_K} e^{-\lambda d(\pi,\sigma)} = \sum_{\pi \in \mathcal{S}_K} e^{-\lambda d(\pi\sigma^{-1},e)} = \sum_{\pi' \in \mathcal{S}_K} e^{-\lambda d(\pi',e)}.$$

The only way the central ranking plays a role in determining $Z(\lambda)$ is through its length K, as it affects the possible range of values $\{0, 1, ..., d_{\max}\}$ for $d(\pi, \sigma)$ and hence

$$Z(\lambda) = \sum_{d=0}^{d_{\max}} n_d e^{-\lambda d},$$

where n_d indicates the number of rankings in $\mathcal{S}_K$ at distance d from σ. It can be easily proved that the normalizing constant is a decreasing bounded function of λ satisfying

$$1 = \lim_{\lambda \to \infty} Z(\lambda) < Z(\lambda) \leq Z(0) = K! \qquad \lambda \geq 0.$$

Figure 1.3 shows the decreasing behaviour of $Z(\lambda)$ in the case of $d = d_K$ and $K = 4$. The determination of $Z(\lambda)$ could be computationally demanding, as it requires the summation over all possible rankings. As stressed in Fligner & Verducci (1986), a possible simplification for its computation relies on the relation of $Z(\lambda)$ with the moment generating function (m.g.f.) $M_D(\cdot)$ of the random distance $D(\cdot, \sigma)$ under the UM on $\mathcal{S}_K$, that is

$$Z(\lambda) = K! \sum_{\pi \in \mathcal{S}_K} e^{-\lambda d(\pi,\sigma)} \frac{1}{K!} = K! \mathbb{E}[e^{-\lambda D}] = K! M_D(-\lambda). \tag{1.2.6}$$

In the wide variety of distances, only some specific ones lead to a closed form expression for $Z(\lambda)$. For this reason when performing a statistical ranked data analysis one should carefully choose an appropriate metric balancing between interpretation purposes, in order to better accommodate the problem at hand, and computational feasibility. In the light of these motivations, in our applications we chose the Kendall distance d_K for the specification of the DB, whereas both d_K and d_C were employed in the inferential procedure for a novel model. These aspects will be better clarified later.

We conclude this section considering some restrictions of the DB. In the cases when σ is supposed to be known, (1.2.5) becomes the element of the one-parameter exponential family originally proposed by Mallows (1957). According to the adopted distance, respectively d_K or d_S^2, such a model is referred to in the literature as *Mallows ϕ- or θ-model*. Actually, Mallows (1957) derived such restricted distance-based models in the attempt to simplify the BSM, setting for its parameters the special form $\rho_{ii'} = (1 + \tanh((i' - i) \log \theta + \log \phi))/2$ and fixing either $\theta = 1$ or $\phi = 1$. These constraints impose that in the ϕ-model the $\rho_{ii'}$'s depend on the sign of $i - i'$, whereas in the θ-model they depend on the difference $i - i'$. More generally, one usually refers to the probability function (1.2.5) as the Mallows model, provided that d is a distance.

1.2.5 Multistage models

In their fundamental work Fligner & Verducci (1988) defined a very general class of probability distributions called *multistage ranking models*. It integrates the idea of the ranking process divided into independent stages, already exploited in the PL, with the presence of a true reference ranking in the population, as assumed in the DB. Postulating the canonical forward (rank-assignment) strategy, any ranking π can be equivalently expressed in terms of a $(K-1)$-dimensional vector $\mathbf{V}(\pi|\sigma) = (V_1(\pi|\sigma), \ldots, V_{K-1}(\pi|\sigma))$ whose generic component $V_t(\pi|\sigma)$ is defined as

$$\begin{aligned} V_t(\pi|\sigma) &= \sigma(\pi^{-1}(t)) - 1 - \sum_{t'<t} I_{[\sigma(\pi^{-1}(t'))<\sigma(\pi^{-1}(t))]} \\ &= \sigma^*(\pi^{-1}(t)) - 1 \qquad\qquad t = 1, \ldots, K-1. \end{aligned} \tag{1.2.7}$$

In (1.2.7) σ^* indicates the reduction of the reference ranking σ on the set of remaining items at stage t, given by $I \setminus \{\pi^{-1}(1), \ldots, \pi^{-1}(t-1)\}$. The vector $\mathbf{V}(\pi|\sigma)$ collects the number of mistakes made by the judge π over the $K-1$ stages w.r.t. the presumed correct ranking σ. Let us make an example to clarify the definition given in (1.2.7) considering $\pi^{-1} = (2,3,4,5,1)$ and $\sigma^{-1} = (4,3,5,1,2)$. This implies $\mathbf{V}(\pi|\sigma) = (4,1,0,0)$. In general the interpretation of entries in $\mathbf{V}(\pi|\sigma)$ is the following: the case $V_t = 0$ states that at stage t the judge π made the correct choice selecting the best item among the available ones, i.e., at that stage her preference matched with the reference σ; on the other hand, when $V_t = v_t > 0$ means that at preference level t the ranker π failed v_t positions, choosing the (v_t+1)-st best item rather than the preferred one by σ among the remaining alternatives. Later we will see that the V_t's are, indeed, the additive components in a measure of global discrepancy between the generic π and the true underlying σ. Exploiting the one-to-one correspondence between π and $\mathbf{V}(\pi|\sigma)$ and assuming that the V_t's are independent, a model setting as

$$\mathbf{P}(\pi|\sigma) = \mathbf{P}(\mathbf{V}(\pi|\sigma)) = \prod_{t=1}^{K-1} \mathbf{P}(V_t(\pi|\sigma) = v_t) \qquad \pi \in \mathcal{S}_K \tag{1.2.8}$$

becomes straightforward. Equation (1.2.8) represents the most general multistage ranking model, indexed by the choice probabilities $\{\mathbf{P}(V_t = v_t) : v_t = 0, \ldots, K - t \text{ and } t = 1, \ldots, K-1\}$. When no constrains are placed on the probability distributions for the V's, Fligner & Verducci (1988) refer to equation (1.2.8) as *free model* (FM). This does not have to be confused with the SM: the independence hypothesis, in fact, leads to a restricted parametric class with $K(K-1)/2$ parameters versus the $K! - 1$ of the SM. We stress that the independence notion employed in (1.2.8) does not refer to the item selection but to the amount of disagreement in making the best choice at each stage w.r.t σ. This aspect reveals the main difference between the PL and the FM: in the former the selection probabilities are indexed by the available alternatives, whereas in the latter they depend only on the stage. Obviously, a probability mass distribution on $\mathcal{S}_K$ induces a probability model on the components of $\mathbf{V}$ and viceversa; in particular, when each V_t is uniformly distributed the FM corresponds to the UM. An important subclass of the FM can be obtained

setting an exponential form for the choice probabilities, that is

$$\mathbf{P}(V_t = v_t|\lambda_t) = \frac{e^{-\lambda_t v_t}}{\sum_{v_t=0}^{K-t} e^{-\lambda_t v_t}} \qquad t = 1, \ldots, K-1. \tag{1.2.9}$$

The restriction (1.2.9) leads to the so-called *ϕ-component model*, governed by the $K-1$ stage-dependent non-negative concentration parameters $\lambda_1, \ldots, \lambda_{K-1}$.

1.2.6 Generalized Mallows model and other DB extensions

In their previous work Fligner & Verducci (1986) had already derived the ϕ-component model starting from a different motivation, specifically illustrating the possibility to extend the DB approach. The starting point is the property of some metrics for rankings, such as the Kendall and the Cayley distance, to be decomposable into the sum of independent components associated to the each single stage of the ranking process. It can be shown, for example, that d_K admits the following multistage formulation

$$d_K(\pi, \sigma) = \sum_{t=1}^{K-1} V_t(\pi|\sigma), \tag{1.2.10}$$

that can be viewed as the decomposition of the global distance in the discrepancies over the ranking stages. Moreover, the decomposition into sums is appealing also because it can lead to a closed form expression for $Z(\lambda)$, provided that the m.g.f. of the single components are simple enough. Hence, Fligner & Verducci (1986) exploited this property of d_K to formulate a parametric generalization of such a distance, applying a non-negative constant to each term as follows

$$d_{\underline{\lambda}}(\pi, \sigma) = \sum_{t=1}^{K-1} \lambda_t V_t(\pi|\sigma), \tag{1.2.11}$$

where $\underline{\lambda} = (\lambda_1, \ldots, \lambda_{K-1})$ is the vector of concentration parameters. This idea is analogous to the introduction of additional stage-dependent quantities in the BM in order to expand the PL family. Fligner & Verducci (1986) further proposed to plug (1.2.11) in (1.2.5), transferring the stagewise construction to the DB and deriving for the latter the following multiparametric generalization

$$P(\pi|\sigma, \underline{\lambda}) = \frac{e^{-\sum_{t=1}^{K-1} \lambda_t V_t(\pi|\sigma)}}{Z(\underline{\lambda})} \qquad \pi \in \mathcal{S}_K, \tag{1.2.12}$$

named *Generalized Mallows model* (GMM). This parametrization allows to assess the closeness to σ in a rank-dependent manner. For example, high values of λ_t for top positions induce a greater probability to observe rankings which agree with σ in the assignment of the first positions. Recognizing in (1.2.12) the product of independent exponential models on the V_t's, the coincidence of the GMM with the ϕ-component model becomes clear. Furthermore, the GMM is associated with the following closed-form expression for the normalizing constant

$$Z(\underline{\lambda}) = \prod_{t=1}^{K-1} Z(\lambda_t) = \prod_{t=1}^{K-1} \sum_{v_t=0}^{K-t} (e^{-\lambda_t})^{v_t} = \prod_{t=1}^{K-1} \frac{1 - e^{-\lambda_t(K-t+1)}}{1 - e^{-\lambda_t}}. \tag{1.2.13}$$

The extension of the DB obtained by the stagewise decomposition of the Cayley distance was called by Fligner & Verducci (1986) *cyclic structure model*, but it is not considered in this review. The equality constraint $\lambda_t = \lambda$ for all $t = 1, \ldots, K-1$ in both generalized models leads directly to the standard DB model with either the $d = d_K$ or the $d = d_C$.

As stressed by Lee & Yu (2010), although model (1.2.12) has been developed to generalize the DB with $d = d_K$, the GMM is not a proper distance-based model. In fact, whereas (1.2.10) is a distance, the same is not true for (1.2.11) because the symmetry property is lost. Inspired by Shieh (1998), Shieh et al. (2000) and Tarsitano (2009), who contributed to define the weighted rank correlation coefficients, Lee & Yu (2010) suggest to employ the weighted versions of the commonly used distances between rankings. These measures permit to evaluate the discrepancy between two rankings by assigning different weights to each rank but, at the same time, retain the desired properties. The derived model is referred to by the authors as *weighted distance-based model* (WDB).

We conclude this section by briefly mentioning the extension of the GMM to address infinite rankings, i.e., the case $K \to \infty$. This situation alludes to all those cases in which the number of items is very high or potentially not completely known, as in the output of search engines or programs for matching physical/biological traits. For this purpose Meilă & Bao (2008) propose the *infinite Generalized Mallows model* (IGM), that they express in a compact form as follows

$$P(\pi|\sigma, \underline{\lambda}) = e^{-\sum_{t=1}^{\infty}(\lambda_t V_t(\pi|\sigma) + \log Z(\lambda_t))} \qquad \pi \in \mathcal{S}_\infty, \tag{1.2.14}$$

where the normalizing constant derives asymptotically from (1.2.13) as follows

$$Z(\lambda_t) = \lim_{K\to\infty} \frac{1 - e^{-\lambda_t(K-t+1)}}{1 - e^{-\lambda_t}} = \frac{1}{1 - e^{-\lambda_t}}.$$

In this case $\mathbf{V}$ defines a vector of independent countably infinite r.v. taking values in $\mathbb{N}$ and the concentration parameters have to be strictly positive in order for (1.2.14) to be a proper probability distribution. In presence of an infinite number of alternatives, it is likely that the ranking process is only partially observed and hence a reasonable assumption is to think of data as subject to a certain mechanism of truncation. Meilă & Bao (2008) discuss the interesting application of the IGM to a situation where K is very large and the recorded rankings are partial top-t sequences. They simply implement the marginalized version of (1.2.14) arrested at the first components of $\mathbf{V}$ actually observed for each subject or, equivalently, at the prefix of $\mathbf{V}$ whose length varies among sample units. The same methodological contribution can be found also in Meilă & Bao (2010).

1.2.7 Insertion Sort Rank data model

Although the models reviewed in the previous sections represent the major and consolidated definitions of stochastic mechanisms on the ranking space, other recent proposals deserve to be mentioned, as the one presented by Biernacki & Jacques (2013). It combines some traditional approaches for rank data modeling, such as the existence of a reference true ordering and the paired comparison scheme, through

the insertion sort algorithm detailed in Knuth (2005). The authors describe a stochastic ranking process which starts from an *initial presentation order* ξ^{-1} and returns each ordered sequence π^{-1} as the result of multiple independent paired comparisons, represented as Bernoulli trials. The probability of success τ is assumed to be constant over all stages of the sorting process and denotes the chance to correctly order a pair of items according to the reference σ^{-1}. These assumptions induce a Binomial form for the following conditional ranking model

$$\mathbf{P}(\pi^{-1}|\xi^{-1},\sigma^{-1},\tau) = \tau^{G(\pi^{-1},\xi^{-1},\sigma^{-1})}(1-\tau)^{A(\pi^{-1},\xi^{-1})-G(\pi^{-1},\xi^{-1},\sigma^{-1})},$$

interpreted as the probability to obtain the generic ordering π^{-1} conditional to the presentation order ξ^{-1}. $A(\pi^{-1},\xi^{-1})$ and $G(\pi^{-1},\xi^{-1},\sigma^{-1})$ indicate respectively the total number of paired comparisons and the number of paired agreements w.r.t. σ^{-1}. The final *insertion sort rank model* $\text{ISR}(\sigma^{-1},\tau)$ is derived integrating out the initial order ξ^{-1} with hypothesized uniform distribution, that is

$$\begin{aligned}\mathbf{P}(\pi^{-1}|\sigma^{-1},\tau) &= \sum_{\xi^{-1}\in\mathcal{S}_K} \mathbf{P}(\pi^{-1}|\xi^{-1},\sigma^{-1},\tau)\mathbf{P}(\xi^{-1}) \\ &= \frac{1}{K!}\sum_{\xi^{-1}\in\mathcal{S}_K} \tau^{G(\pi^{-1},\xi^{-1},\sigma^{-1})}(1-\tau)^{A(\pi^{-1},\xi^{-1})-G(\pi^{-1},\xi^{-1},\sigma^{-1})}.\end{aligned}$$

The reader is referred to Biernacki & Jacques (2013) for further details on the formal properties of the $\text{ISR}(\sigma^{-1},\tau)$ and the implemented inferential procedures.

1.3 Novel extension of the Plackett-Luce model

In this section we introduce an original proposal which generalizes the standard PL. Multistage ranking models previously reviewed implicitly suppose that preferences are expressed with the canonical forward order, proceeding from the assignment of the first rank up to the last one. This is just a conventional assumption and other reference orders can be contemplated but, to our knowledge, this aspect has not been addressed in the literature. Indeed, even the individual experience in choice problems suggests the plausibility of alternative paths for the ranking elicitation. For example, one can think of situations where the judge has a clearer perception about her most- and least-liked items but only a vaguer idea relative to middle ranks; alternatively the ranker could build up her best alternatives following an exclusion process starting with the final position, which would be better described by a backward model. Besides such motivations aimed at characterizing typical behaviors in real choice/selection problems, we are also interested in developing a more flexible tool in order to improve the description of phenomena observed in the form of ordered data. All these intuitive and practical instances make the forward hypothesis too restrictive when approaching a flexible inferential analysis of a ranking data set. Hence, we propose to extend the PL in the following way: rather than fixing *a priori* the stepwise order leading the judge to her final ranked sequence, we represent it with a specific free model parameter $\rho \in \mathcal{S}_K$ and let data guide inference about the reference order followed in the rank assignment scheme. Such an approach would also alleviate the asymmetry toward ranks assigned at the

extreme (the first and the last) stages of the ranking procedure, which affects the PL. The reference order $\rho = (\rho(1), \dots, \rho(K))$ is the result of a bijection between the stage set S and the rank set R, i.e.,

$$\rho : S \to R,$$

where the entry $\rho(t)$ indicates the rank attributed at the t-th stage of the ranking process. Hence, ρ identifies a discrete parameter taking values in $\mathcal{S}_K$. Once alternatives schemes for the ranking elicitation are contemplated, the sequence π^{-1} no longer coincides with the actual item selections over the stages. In order to reconstruct this information, the composition of the ordering π^{-1} with the reference order ρ has to be considered, yielding the sequence

$$\eta^{-1} = \pi^{-1}\rho$$

which lists the items chosen at each stage. This means that $\eta^{-1}(t) = \pi^{-1}(\rho(t))$ is the item chosen at step t and receiving rank $\rho(t)$. The probability of a random ordering under the *Extended Plackett-Luce model* can be written as

$$\mathbf{P}_{\text{EPL}}(\pi^{-1}|\rho, \underline{p}) = \mathbf{P}_{\text{PL}}(\pi^{-1}\rho|\underline{p}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}}{\sum_{\nu=t}^{K} p_{\pi^{-1}(\rho(\nu))}} \qquad \pi^{-1} \in \mathcal{S}_K, \tag{1.3.1}$$

where the additional discrete parameter ρ acts directly by composition on the right-hand side of the generated outcome of a standard PL. The composition determines the rearrangement of components in π^{-1} and reveals the actual item selections over the stages. Hereafter we will shortly refer to (1.3.1) as EPL(ρ,$\underline{p}$). The vector $\underline{p}$ still denotes the support parameters, with the probabilities for each item to be selected at the first stage and to be ranked in the position given by the first entry in ρ. Obviously, the standard PL is a special case of the EPL, obtained setting ρ equal to the identity permutation e. Similarly, when $\rho = (K+1) - e$ one has the backward PL.

Let us make a toy example elucidating the definition and the notation of the EPL given in (1.3.1). Suppose we have fixed a parameter point $(\rho, \underline{p})$ so that $\underline{p} = (0.1, 0.2, 0.3, 0.4)$ and the entries of the reference order $\rho = (1, 4, 2, 3)$ correspond to the following alternating selection scheme: at the first stage the judge expresses her best preference ($\rho(1) = 1$), at the second stage she chooses her least-liked item ($\rho(2) = 4$) and finally at the third and forth stage she attributes respectively the second ($\rho(3) = 2$) and the third ($\rho(4) = 3$) position. Let $\pi^{-1} = (4, 3, 1, 2)$ be the ordering of interest for which we want to compute the probability mass under the specified EPL. The EPL retains the same stagewise ranking construction of the PL but allows the rank attribution order to be different from the ordinary best-to-worst path. The EPL postulates that the probability associated to π^{-1} is equivalent to the probability under the PL of sequentially selecting item 4 at the first stage, item 2 at the second stage, item 3 at the third stage and item 1 at the last stage, as indicated by the composition

$$\begin{aligned} \pi^{-1}\rho &= (\pi^{-1}(\rho(1)), \pi^{-1}(\rho(2)), \pi^{-1}(\rho(3)), \pi^{-1}(\rho(4))) \\ &= (\pi^{-1}(1), \pi^{-1}(4), \pi^{-1}(2), \pi^{-1}(3)) = (4, 2, 3, 1). \end{aligned}$$

Hence, the probability can be computed as

$$\begin{aligned}\mathbf{P}_{\mathrm{EPL}}((4,3,1,2)|\rho,\underline{p}) &= \mathbf{P}_{\mathrm{PL}}((4,2,3,1)|(0.1,0.2,0.3,0.4)) \\ &= \frac{0.4}{1}\cdot\frac{0.2}{0.1+0.2+0.3}\cdot\frac{0.3}{0.1+0.3}\cdot 1 = 0.1.\end{aligned}$$

Note that, in order to interpret the order of preferences from $\pi^{-1}\rho$, one needs to refer to the reference order ρ.

From a theoretical point of view, the novel EPL is a proper generalization of the original PL if and only if such a new class covers a wider portion of the SM, i.e., it allows to describe additional probability functions that cannot be derived from the PL with any parameter specification. In other words, one should give a formal proof concerning the existence of ranking distributions (hopefully more than one) in the new EPL which do not belong to the standard PL family. Such a proof is given in the next section.

1.4 Relation between the novel EPL and the PL

We prove here the presence of distributions on orderings in the novel EPL family which are not members of the canonical PL family. For this purpose, we recall that the PL implies the *independence of irrelevant alternatives* (IIA), stating that the probability ratio of selecting an item over another is unaffected by the preferences towards the other alternatives in the choice set (see Luce (1959)). Equivalently, one can say that in a PL the choice probability ratio between two items is constant over all stages as long as such alternatives are both still available. In the case $K = 3$ the IIA property translates into the following set of conditions on the probabilities $q_{\pi^{-1}} = \mathbf{P}(\pi^{-1})$ of each possible ordering

$$\begin{cases} \dfrac{q_{(1,2,3)}}{q_{(1,3,2)}} = \dfrac{q_{(2,1,3)}+q_{(2,3,1)}}{q_{(3,1,2)}+q_{(3,2,1)}}, \\ \dfrac{q_{(2,1,3)}}{q_{(2,3,1)}} = \dfrac{q_{(1,2,3)}+q_{(1,3,2)}}{q_{(3,1,2)}+q_{(3,2,1)}}, \\ \dfrac{q_{(3,1,2)}}{q_{(3,2,1)}} = \dfrac{q_{(1,2,3)}+q_{(1,3,2)}}{q_{(2,1,3)}+q_{(2,3,1)}}. \end{cases} \tag{1.4.1}$$

The first condition, for example, imposes the equality between the probability ratio to prefer item 2 over item 3 at the second stage (left-hand side) and the analogous probability ratio for the first stage (right-hand side). The remaining equalities have the same meaning but clearly refer to the other possible paired comparisons. The equations in (1.4.1) have to be simultaneously satisfied for a generic ranking distribution to belong to the forward PL. Now, let us consider the EPL with parameter $\rho = (2,1,3)$ and generic support parameter vector $\underline{p}$. The induced probability distribution on the random orderings is given by

$$\begin{pmatrix} q_{(1,2,3)} & q_{(1,3,2)} & q_{(2,1,3)} & q_{(2,3,1)} & q_{(3,1,2)} & q_{(3,2,1)} \\ \dfrac{p_2p_1}{1-p_2} & \dfrac{p_3p_1}{1-p_3} & \dfrac{p_1p_2}{1-p_1} & \dfrac{p_3p_2}{1-p_3} & \dfrac{p_1p_3}{1-p_1} & \dfrac{p_2p_3}{1-p_2} \end{pmatrix}. \tag{1.4.2}$$

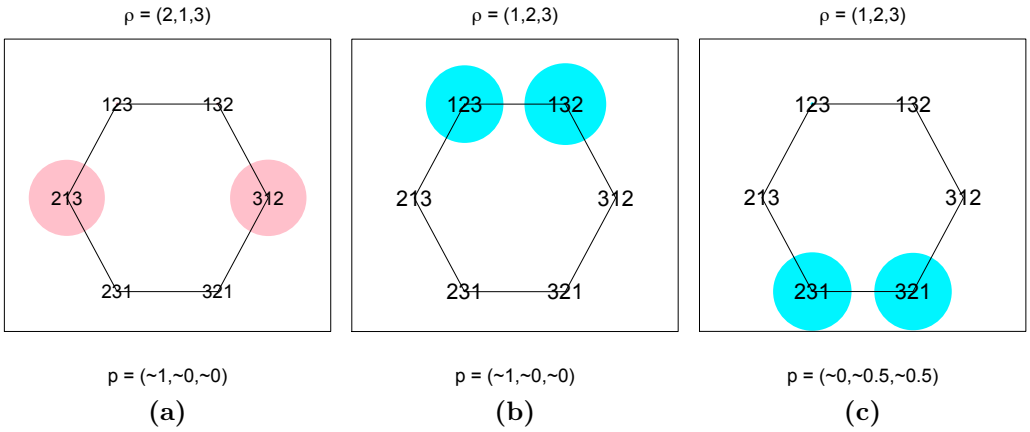

Figure 1.4. Examples of EPL (left) and PL (center and right) distribution functions in the case $K = 3$. Sequences at the vertices indicate orderings.

Substituting (1.4.2) in (1.4.1) and solving w.r.t. $\underline{p}$ one obtains as unique solution $\underline{p} = (1/3, 1/3, 1/3)$, meaning that the two model classes can share only the UM. This formally shows what has been hinted at in Fligner & Verducci (1988) on the possibility to define new ranking models relaxing the forward hypothesis. To give an intuition about the types of ranking distributions that are not covered by the traditional PL, let us consider the EPL with parameter configuration $\rho = (2, 1, 3)$ and $\underline{p} = (1 - 2\epsilon, \epsilon, \epsilon)$ where $\epsilon \to 0$. The corresponding distribution over the six possible orderings has two equally supported modes on the sequences with item 1 ranked second, which capture almost the total probability mass as shown in Figure 1.4(a). This represents a distribution that cannot be obtained with any parameter specification from the forward PL. In fact, the suitable calibration of the support parameters in the PL can lead only to degenerate marginal choices of item 1 for the first and the last rank, see Figures 1.4(b) and 1.4(c). Therefore, the introduction of the extra parameter ρ running over the permutation space allows to overcome this asymmetry among ranks.

1.5 Finite mixture modeling for ranked data

One of the formal properties satisfied by the DB with $d = d_K$ is *strong unimodality*, meaning that the distribution is uniquely maximized by σ and the probability does not increase when the distance from the modal sequence becomes higher (see Marden (1995)). This representation implies a homogeneous population clustered around a single modal sequence, where the degree of agreement w.r.t. the true ranking σ is controlled by the concentration parameter λ. Strong unimodality is expected to be violated in real data especially when the sample composition is heterogeneous w.r.t. factors related to the ranking elicitation. In this case the unimodality assumption should be relaxed in favor of a multimodal characterization of the sampling distribution. A well-established statistical tool to address inference in the presence of unobserved heterogeneity is given by the finite mixture approach. A *finite mixture model* assumes that the population consists of a finite number G of subpopulations.

In this setting the probability of observing the ranking π_s for the s-th unit is

$$f(\pi_s) = \sum_{g=1}^{G} \omega_g f_g(\pi_s) \qquad \pi_s \in \mathcal{S}_K,$$

where $f_g(\cdot)$ denotes the g-th *component* of the mixture, i.e., the statistical distribution of data within the g-th group, and ω_g is the probability for the s-th observation to belong to the g-th group. The membership probabilities $\underline{\omega} = (\omega_1, \ldots, \omega_G)$ are usually termed *weights* of the mixture. Mixture components are often modeled with members of the same parametric family, that is $f_g(\cdot) = f(\cdot|\theta_g) \in \{f(\cdot|\theta) : \theta \in \Theta\}$ for all $g = 1, \ldots, G$, identified by the group-specific parameters θ_g. For a more extensive introduction to finite mixture models the reader can refer to McLachlan & Peel (2000).

Beside a clustering-oriented analysis, aimed at recognizing differential patterns in the data, mixture modeling is also motivated as effective tool to describe less structured populations, in alternative to a nonparametric approach. In this perspective the generalization to the finite mixture makes all generative ranking models previously reviewed more flexible and enlarges significantly their applicability. For this reason, after the early seminal papers by Croon (1989) and Croon & Luijkx (1993), an increasing recourse to the mixture framework in the more recent literature can be highlighted. For example, Murphy & Martin (2003) analyzed the popular 1980 APA (American Psychological Association) presidential election data set (specifically the sub-data set of complete rankings) fitting a mixture of distance-based models. They aimed at inquiring voters' orientation towards candidates within the electorate, assessing the possible adequacy to incorporate a noise component (UM) in the mixture. Such a component, in fact, could collect outliers and/or observations characterized by untypical preference profiles, with a possible final improvement of model fitting. A similar approach has been adopted in Lee & Yu (2012), who estimated a mixture of WDB from the German sample data set, and in other preference studies in combination with an increasing interest towards more general methods to analyze heterogeneous (partial+full) ranking data. In this context, parametric models based on a stagewise ranking construction are attractive because they allow a straightforward treatment of mixed-type information once their marginal form is considered. Although multistage models apply in a natural way to accomodate inference based simultaneously on full and top-t partial rankings, contributions are not limited only to this type of models. Among the applications to partial rankings one can mention Busse et al. (2007), who performed a model-based clustering of the entire APA data set applying a maximum entropy approach for the DB with $d = d_K$, and the nonparametric clustering algorithm, named Exponential-Blurring-Mean-Shift, for the IGM implemented by Meilă & Bao (2008). Gormley & Murphy (2006) fitted a mixture of PL to the 2000 CAO (Central Applicant Office) data set to investigate motivations of Irish college applicants in their degree course choice, whereas Gormley & Murphy (2008a) employed a mixtures of both PL and BM to infer the structure of the Irish political electorate and characterize voting blocks. In more recent works the same authors attempted to extend the mixture framework in different directions. Gormley & Murphy (2008b) and Gormley & Murphy (2010), for example, contributed to the estimation of the mixture of experts models (MOE)

for ranked data. The MOE approach was formalized in the machine learning literature by Jacobs et al. (1991) and involves as special case the *concomitant-variable latent class model*, proposed in the statistical methodology by Dayton & Macready (1988). It accommodates for the introduction of individual covariates in the finite mixture through the generalized linear model theory. The covariates can be applied to either mixing proportions or component paramaters (or both), leading to different MOE specifications. Thus, the MOE is flexible tool allowing to explore simultaneously the groups configuration and the impact of social/economic factors on preferences.

In Section 3.3 will verify the utility of alternative flexible classes of ranking models in a real application, considering data from the LFPD bioassay experiment. For the analysis of the LFPD data set we will implement different mixture models, adopting as mixture components elements from the following parametric families:

- DB with $d = d_K$,
- PL with known forward and backward reference order,
- BM,
- our novel EPL.

DB and PL represent two of the most frequently used distributions for inferring ranking data and both parameterizations allow a clear interpretation. In the former case, the central ranking summarizes the overall profile in assessing the orderings of the items, whereas the concentration parameter expresses how representative the modal ranking is. In the latter case, a higher value of the item support parameter implies a greater probability for that item to be preferred at each selection level. The implementation of the BM mixture facilitates the investigation of how the BM and the EPL compete in terms of flexibility, although they generalize the PL in totally different directions. This analysis and related inferential details discussed in Section 2.2 are the main contribution to ranked data modeling proposed in Mollica & Tardella (2014).

In Chapter 2 we will rely on maximum likelihood estimation, typically achieved with iterative optimization procedures. Since the Bayesian paradigm provides a more general inferential framework with the important advantage to account for estimation uncertainty in a more straightforward manner, contributions to Bayesian ranked data modeling are reviewed separately in Chapter 5, with a special focus on the PL and a further proposal.

Chapter 2

Maximum likelihood inference for ranking models

In this chapter the attention is devoted to the inferential aspects for the parameter estimation in the maximum likelihood approach, starting with the simpler case of ranked data models for a homogenous population and subsequently generalizing it to the finite mixture setting with G components. For the stagewise models we limit ourselves to give details only for the novel EPL distribution because, as mentioned in Section 1.3, the conventional forward PL is a reduction of the wider EPL family. It follows that the estimation procedure for the mixture of PL can be easily derived from the mixture of EPL with all known reference orders $\rho_g = e$ for $g = 1, \ldots, G$. However, explicit estimation formulas for this special case can be found in Gormley & Murphy (2006), whereas inference concerning the mixture of BM is detailed in Gormley & Murphy (2008a). We begin with the MLE of the DB considering, without too much loss of generality, the case where the Kendall distance is assumed in the model specification.

2.1 MLE of the mixture of distance-based models

Let $\underline{\pi} = \{\pi_1, \ldots, \pi_N\}$ be a random sample drawn from the following homogenous population

$$\pi_1, \ldots, \pi_N | \sigma, \lambda \overset{i.i.d.}{\sim} \text{DB}(\sigma, \lambda).$$

The corresponding log-likelihood turns out to be

$$l(\sigma, \lambda) = -\Big(\lambda \sum_{s=1}^{N} d_K(\pi_s, \sigma) + N \log Z(\lambda)\Big), \tag{2.1.1}$$

which is jointly optimized with a two-step procedure. Since for any value of lambda the log-likelihood is a strictly decreasing function of the sum of the empirical distances from the central sequence σ, in the first step one maximizes the function (2.1.1) w.r.t. σ as follows

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{s=1}^{N} d_K(\pi_s, \sigma).$$

In the second step one uses the profile log-likelihood $l_{\text{prof}}(\lambda) = l(\lambda, \hat{\sigma})$ to get the MLE as follows

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}_0^+}{\arg\max}\ l_{\text{prof}}(\lambda).$$

Borrowing the property (1.2.6) one can compute the $\hat{\lambda}$ as the solution of

$$\bar{d}_K = -\frac{Z'(\lambda)}{Z(\lambda)} \Longleftrightarrow \bar{d}_K = \frac{M'_{D_K}(-\lambda)}{M_{D_K}(-\lambda)} \Longleftrightarrow \bar{d}_K = \frac{d}{dt} \log M_{D_K}(t)\Big|_{t=-\lambda},$$

where $\bar{d}_K$ is equal to $\sum_{s=1}^{N} d_K(\pi_s, \hat{\sigma})/N$. In rank aggregation theory, which aims at synthesizing a certain set of rankings in a single overall ordered sequence, $\hat{\sigma}$ is known as the *Kemeny ranking* and is NP-hard to compute. For large values of K, in fact, the exhaustive search in $\mathcal{S}_K$ needed to get $\hat{\sigma}$ becomes unfeasible. Fligner & Verducci (1988) suggest to alternate sequentially the approximate maximization over σ through a local search and the solution of the estimating equation for λ. In the local search one minimizes the sum of the sample distances over the subset of rankings adjacent to the current estimate. Notice that adjacency can be considered in terms of a metric which need not be the same used in the model specification (1.2.5). The *Borda ranking* $\overline{\overline{\pi}} = (\overline{\overline{\pi}}(1), \ldots, \overline{\overline{\pi}}(K))$, whose generic component is defined from the average rank vector as

$$\overline{\overline{\pi}}(i) = \text{rank}(\overline{\pi}(i))$$

in $\{\overline{\pi}(1), \ldots, \overline{\pi}(K)\}$, is often used to initialize σ. Other reasonable starting values can be deduced by the wide variety of rules developed in the rank aggregation literature. For an in-depth simulation-based analysis comparing several methods to address the computational complexity of the Kemeny ranking problem, the recent work by Ali & Meilă (2012) is recommended. For the estimation of λ we recall that the DB is an element of the exponential family when σ is fixed, implying that the following relations for the generic random distance $D(\cdot, \sigma)$ hold

$$M_{\lambda,D}(t) = M_D(t - \lambda)/M_D(-\lambda)$$

and

$$\mathbb{E}_{\sigma,\lambda}[D] = \frac{d}{dt} \log M_D(t)\Big|_{t=-\lambda} \qquad \mathbb{VAR}_{\sigma,\lambda}[D] = \frac{d^2}{dt^2} \log M_D(t)\Big|_{t=-\lambda},$$

where $M_{\lambda,D}(t)$, $\mathbb{E}_{\sigma,\lambda}[D]$ and $\mathbb{VAR}_{\sigma,\lambda}[D]$ denote the m.g.f., the expectation and the variance of $D(\cdot, \sigma)$ w.r.t. (1.2.5) (Fligner & Verducci, 1986). Note that we can drop σ in the subscript notation to stress the independence of the distribution of D on the central ranking, due to label invariance of the distance. Thus, the MLE of λ satisfies

$$\bar{d}_K = \mathbb{E}_{\hat{\lambda}}[D_K],$$

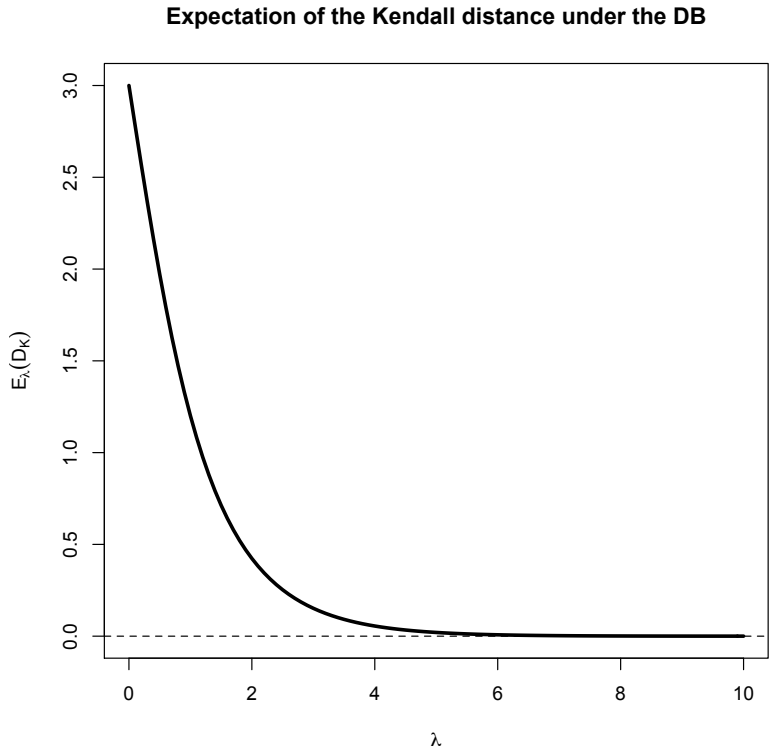

Figure 2.1. Expectation of the Kendall distance D_K under the distance-based model for $K = 4$ as function of the concentration parameter λ.

i.e., it corresponds to the value of the concentration that matches the expectation of D with the sample average distance. In the specific case $D = D_K$, one can write

$$\begin{aligned} M_{D_K}(t) &= \mathbb{E}\left[e^{tD_K(\pi,\sigma)}\right] = \mathbb{E}\left[e^{t\sum_{j=1}^{K-1} V_j(\pi|\sigma)}\right] = \mathbb{E}\left[\prod_{j=1}^{K-1} e^{tV_j(\pi|\sigma)}\right] \\ &\overset{i.}{=} \prod_{j=1}^{K-1} \mathbb{E}\left[e^{tV_j(\pi|\sigma)}\right] = \prod_{j=1}^{K-1}\left(\frac{1}{(K-j+1)}\sum_{v_j=0}^{K-j} e^{tv_j}\right) \\ &= \frac{1}{K!}\prod_{j=1}^{K-1}\frac{1-e^{t(K-j+1)}}{1-e^t} = \frac{(1-e^t)^{-K}}{K!}\prod_{j=1}^{K}\left(1-e^{t(K-j+1)}\right) \end{aligned}$$

and

$$\frac{d}{dt}\log M_{D_K}(t) = \frac{Ke^t}{1-e^t} - \sum_{j=1}^{K}(K-j+1)\frac{e^{t(K-j+1)}}{1-e^{t(K-j+1)}} = \frac{Ke^t}{1-e^t} - \sum_{j=1}^{K}\frac{je^{jt}}{1-e^{jt}},$$

hence

$$\mathbb{E}_\lambda[D_K] = \frac{Ke^{-\lambda}}{1-e^{-\lambda}} - \sum_{j=1}^{K}\frac{je^{-j\lambda}}{1-e^{-j\lambda}}.$$

As plotted in Figure 2.1, $\mathbb{E}_\lambda[D_K]$ is a monotone decreasing function in λ. Its shape is similar to the one observed for $Z(\lambda)$ and also in this case we have a bounded function.

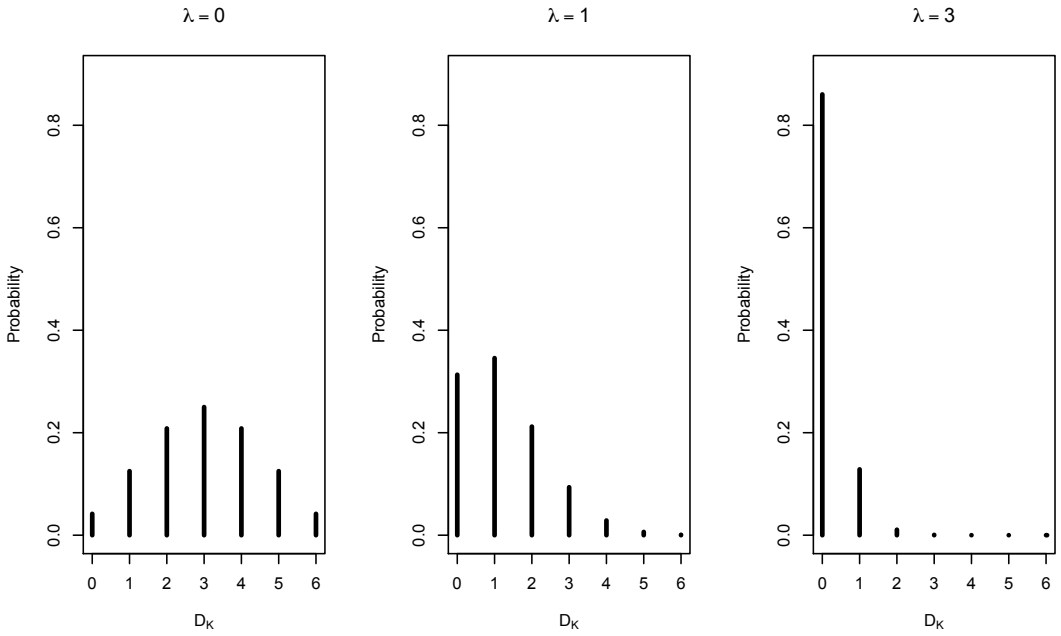

Figure 2.2. Distribution of the Kendall distance D_K under the distance-based model for $K = 4$ and different values of the concentration parameter: $\lambda = 0$ (left), $\lambda = 1$ (center) and $\lambda = 3$ (right).

In particular, under the UM $\mathbb{E}_0[D_K] = K(K-1)/4$ because of the symmetric distribution of D_K over the discrete support $\{0, 1, \ldots, K(K-1)/2\}$. As λ increases, the probability corresponding to lower values of D_K becomes higher and $\mathbb{E}_\lambda[D_K]$ decreases towards zero for $\lambda \to \infty$. Examples of distributions of D_K for different values of λ are shown in Figure 2.2. An approximation of $\hat{\lambda}$ can be easily derived using the table provided by Feigin & Cohen (1978) (limited up to $K = 10$ and expressed in the parametrization $\theta = e^{-\lambda}$) or by a line search algorithm.

Now we summarize the fundamental steps to derive the MLE for a mixture of DB with $d = d_K$. We basically reproduce the algorithm described in Murphy & Martin (2003). Let $\underline{z}_s = (z_{s1}, \ldots, z_{sG})$ be the latent variables indicating the individual component membership such that

$$z_{sg} = \begin{cases} 1 & \text{if the } s\text{-th unit belongs to the } g\text{-th group,} \\ 0 & \text{otherwise.} \end{cases}$$

From (1.2.5) it follows that the complete log-likelihood can be expressed as

$$\begin{aligned} l_c(\underline{\sigma}, \underline{\lambda}, \underline{\omega}, \underline{z}) &= \log \mathbf{P}(\underline{\pi}, \underline{z}|\underline{\sigma}, \underline{\lambda}, \underline{\omega}) = \log \prod_{s=1}^{N} \mathbf{P}(\pi_s, \underline{z}_s|\underline{\sigma}, \underline{\lambda}, \underline{\omega}) \\ &= \log \prod_{s=1}^{N} \mathbf{P}(\pi_s|\underline{z}_s, \underline{\sigma}, \underline{\lambda}) \mathbf{P}(\underline{z}_s|\underline{\omega}) \\ &= \sum_{s=1}^{N} \sum_{g=1}^{G} \log\left(\omega_g \frac{e^{-\lambda_g d_K(\pi_s, \sigma_g)}}{Z(\lambda_g)}\right)^{z_{sg}} \\ &= \sum_{s=1}^{N} \sum_{g=1}^{G} z_{sg}\big(\log \omega_g - \lambda_g d_K(\pi_s, \sigma_g) - \log Z(\lambda_g)\big), \end{aligned}$$

where $\underline{\omega}$ and $\underline{\lambda}$ are vectors representing, respectively, the group membership probabilities and the component-specific concentration parameters, whereas $\underline{\sigma}$ is a $G \times K$ matrix, whose rows indicate the central rankings of the mixture components. In order to derive parameter estimates, the EM algorithm can be implemented (Dempster et al., 1977). It represents the major scheme to address the inferential analysis in the presence of missing data. For the present model the EM algorithm consists of the following steps:

Initialization: set initial values $\underline{\sigma}^{(0)}, \underline{\lambda}^{(0)}, \underline{\omega}^{(0)}$ for the parameters to be estimated;

at iteration $l+1$ compute

E-step: for $s = 1, \ldots, N$ and $g = 1, \ldots, G$

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} \mathbf{P}_{\mathrm{DB}}(\pi_s|\sigma_g^{(l)}, \lambda_g^{(l)})}{\sum_{g'=1}^{G} \omega_{g'}^{(l)} \mathbf{P}_{\mathrm{DB}}(\pi_s|\sigma_{g'}^{(l)}, \lambda_{g'}^{(l)})},$$

which represent the current estimates of the posterior membership probabilities;

M-step: for $g = 1, \ldots, G$

$$\begin{aligned} \omega_g^{(l+1)} &= \sum_{s=1}^{N} \frac{\hat{z}_{sg}^{(l+1)}}{N}, \\ \sigma_g^{(l+1)} &= \arg\min_{\sigma_g} \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} d(\pi_s, \sigma_g) \end{aligned}$$

and determine $\lambda_g^{(l+1)}$ as the solution of

$$\frac{K e^{-\lambda_g}}{1 - e^{-\lambda_g}} - \sum_{j=1}^{K} \frac{j e^{-j\lambda_g}}{1 - e^{-j\lambda_g}} = \frac{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} d(\pi_s, \sigma_g^{(l+1)})}{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}.$$

Obviously, setting $G = 1$ these inferential equations reduce to the estimates previously described for the homogenous model.

2.2 MLE of the mixture of Extended Plackett-Luce models

The derivation of the MLE for the EPL is less straightforward and requires a suitable adjustment for the maximization step. Assuming the EPL(ρ,$\underline{p}$) as the sampling distribution of the observed orderings, the log-likelihood has the following expression

$$\begin{aligned} l(\rho, \underline{p}) &= \sum_{s=1}^{N}\sum_{t=1}^{K} \log \frac{p_{\pi_s^{-1}(\rho(t))}}{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}} \\ &= \sum_{s=1}^{N}\sum_{t=1}^{K} \left(\log p_{\pi_s^{-1}(\rho(t))} - \log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))} \right) \\ &= N \sum_{i=1}^{K} \log p_i - \sum_{s=1}^{N}\sum_{t=1}^{K} \log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}. \end{aligned} \tag{2.2.1}$$

Note that the direct maximization of the log-likelihood w.r.t. the p's is made arduous by the presence of the annoying term $\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}$. Therefore, we derive the estimation formula for the support parameters borrowing the approach in Hunter (2004) based on the *Minorization/Maximization* (MM) algorithm. This iterative optimization method is reviewed in its general form by Lange et al. (2000) and Hunter & Lange (2004), whereas Hunter (2004) discusses the specific application of the MM algorithm for the PL estimation. The basic idea consists in performing the optimization step for the p's on a surrogate objective function rather than on (2.2.1). The surrogate is obtained by exploiting the strict convexity of $-\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}$ and in particular the supporting hyperplane property for convex functions. From Taylor's linear expansion, in fact, one has

$$-\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))} \geq 1 - \log \sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho(\nu))} - \frac{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}}{\sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho(\nu))}}.$$

For optimization purposes the additive term $1 - \log \sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho(\nu))}$ not depending on $\underline{p}$ can be disregarded. With the suitable substitutions the minorizing auxiliary objective function can be written as

$$q = N \sum_{i=1}^{K} \log p_i - \sum_{s=1}^{N}\sum_{t=1}^{K} \frac{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}}{\sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho(\nu))}}. \tag{2.2.2}$$

As emphasized by Hunter (2004), the advantage of optimizing the more tractable (2.2.2) in place of (2.2.1) relies on the separation of the support parameters. Furthermore, in Hunter (2004) it is proved that the iterative maximization of q returns a sequence $\underline{p}^{(1)}, \underline{p}^{(2)}, \ldots$ which converges at least to a local maximum of the original objective function. Thus, differentiating the surrogate w.r.t. each p_i one has

$$\frac{\partial q}{\partial p_i} = \frac{N}{p_i} - \sum_{s=1}^{N}\sum_{t=1}^{K} \frac{\delta^{(l)}_{sti}}{\sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho(\nu))}} \tag{2.2.3}$$

and equating the partial derivatives (2.2.3) to zero, at the current iteration one obtains the following updating rule for the support parameters

$$p_i^{(l+1)} = \frac{N}{\sum_{s=1}^{N}\sum_{t=1}^{K}\dfrac{\delta_{sti}^{(l)}}{\sum_{\nu=t}^{K} p^{(l)}_{\pi_s^{-1}(\rho^{(l)}(\nu))}}} \qquad i = 1, \ldots, K.$$

The binary indicator

$$\delta_{sti}^{(l)} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(\rho^{(l)}(t)), \ldots, \pi_s^{-1}(\rho^{(l)}(K))\}, \\ 0 & \text{otherwise} \end{cases}$$

indicates the event that item i is still available at stage t for the subject s or, equivalently, that is not selected by unit s before stage t. Notice that the binary array has a superscript because of the dependence on the $\rho = \rho^{(l)}$ available at the current iteration. In the MLE of the PL, instead, this array is not subject to update (Gormley & Murphy, 2006).

Using the original log-likelihood, the update of the reference order parameter is derived as follows

$$\rho^{(l+1)} = \arg\min_{\rho} \sum_{s=1}^{N}\sum_{t=1}^{K} \log \sum_{\nu=t}^{K} p^{(l+1)}_{\pi_s^{-1}(\rho(\nu))}. \tag{2.2.4}$$

Solving (2.2.4) with a global search in $\mathcal{S}_K$ is prohibitive when K is large. So, we implemented a local search similarly to the method suggested by Busse et al. (2007) and Lee & Yu (2010), constraining the optimization within a fixed distance from the current estimate of the reference order $\rho^{(l)}$. It may be interesting to evaluate the sensitivity of the algorithm w.r.t. the choice of a particular distance in the local search step. In Section 3.3 we will perform such sensitivity analysis focusing only on the Kendall and Cayley distance and compare the corresponding estimation performances.

Now we relax the hypothesis of homogeneous population and consider a more flexible mixture model with EPL components, discussing the related inferential issues. Augmenting data with the missing individual group membership variables $\underline{z}_s = (z_{s1}, \ldots, z_{sG})$, one obtains the following expression for the complete log-likelihood

$$\begin{aligned} l_c(\underline{\rho}, \underline{p}, \underline{\omega}, \underline{z}) &= \log \prod_{s=1}^{N}\prod_{g=1}^{G} \left(\omega_g \prod_{t=1}^{K} \frac{p_{g\pi_s^{-1}(\rho_g(t))}}{\sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\rho_g(\nu))}} \right)^{z_{sg}} \\ &= \sum_{s=1}^{N}\sum_{g=1}^{G} z_{sg} \left(\log \omega_g + \sum_{i=1}^{K} \log p_{gi} - \sum_{t=1}^{K} \log \sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\rho_g(\nu))} \right). \end{aligned}$$

In the EM algorithm the maximization problem is transferred to the the expectation of the l_c w.r.t. the posterior distribution of the latent variables $\underline{z}$ represented by $\hat{\underline{z}}$, that is

$$\begin{aligned} Q((\underline{\rho}, \underline{p}, \underline{\omega}), (\underline{\rho}^*, \underline{p}^*, \underline{\omega}^*)) &= \mathbb{E}[l_c | \underline{\pi}^{-1}, \underline{\rho}^*, \underline{p}^*, \underline{\omega}^*] \\ &= \sum_{s=1}^{N}\sum_{g=1}^{G} \hat{z}_{sg} \left(\log \omega_g + \sum_{i=1}^{K} \log p_{gi} - \sum_{t=1}^{K} \log \sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\rho_g(\nu))} \right), \end{aligned}$$

where for $s = 1, \ldots, N$ and $g = 1, \ldots, G$

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} \mathbf{P}_{\mathrm{EPL}}(\pi_s^{-1} | \rho_g^{(l)}, \underline{p}_g^{(l)})}{\sum_{g'=1}^{G} \omega_{g'}^{(l)} \mathbf{P}_{\mathrm{EPL}}(\pi_s^{-1} | \rho_{g'}^{(l)}, \underline{p}_{g'}^{(l)})}.$$

Similarly to Gormley & Murphy (2006) we combined the EM with the MM algorithm into a hybrid version of the former, called EMM algorithm, using the following minorizing surrogate function q in place of the original objective

$$Q((\underline{\rho}, \underline{p}, \underline{\omega}), (\underline{\rho}^*, \underline{p}^*, \underline{\omega}^*)) \geq q$$
$$= \sum_{s=1}^{N} \sum_{g=1}^{G} \hat{z}_{sg} \sum_{i=1}^{K} \log p_{gi} - \sum_{s=1}^{N} \sum_{g=1}^{G} \hat{z}_{sg} \sum_{t=1}^{K} \frac{\sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\rho_g(\nu))}}{\sum_{\nu=t}^{K} p^{(l)}_{g\pi_s^{-1}(\rho_g(\nu))}}.$$

The maximization of q leads to the updating rule for p_{gi} at the current iteration given by

$$p_{gi}^{(l+1)} = \frac{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} \sum_{t=1}^{K} \frac{\delta_{stig}^{(l)}}{\sum_{\nu=t}^{K} p^{(l)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}}}$$

for $g = 1, \ldots, G$ and $i = 1, \ldots, K$, with

$$\delta_{stig}^{(l)} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(\rho_g^{(l)}(t)), \ldots, \pi_s^{-1}(\rho_g^{(l)}(K))\}, \\ 0 & \text{otherwise,} \end{cases}$$

indicating if, under the group-specific reference order ρ_g, the unit s did not choose the i-th item before stage t, and hence if, at that step, it still belongs to the set of available alternatives or not. The estimate for the reference orders in each subgroup is the solution of the optimization problem given, for $g = 1, \ldots, G$, by

$$\rho_g^{(l+1)} = \arg\min_{\rho_g} \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} \sum_{t=1}^{K} \log \sum_{\nu=t}^{K} p^{(l+1)}_{g\pi_s^{-1}(\rho_g(\nu))},$$

which can be solved locally with ρ ranging in a suitable neighborhood of the current reference order, as in the case $G = 1$. The M-step ends with the traditional update of the mixture weights

$$\omega_g^{(l+1)} = \frac{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}{N} \qquad g = 1, \ldots, G,$$

computed as the posterior proportions of sample units belonging to each group. Finally, in order to address the issue of local maxima, we run the algorithm with a suitably large number of different starting values.

2.3 Algorithm convergence and model selection

As detailed in the previous sections, we conducted MLE relying on the EM algorithm and on a hybrid version thereof. For this purpose, we developed a suite of functions

written in R language (R Core Team, 2012). In these estimation procedures the log-likelihood is iteratively maximized until a suitable convergence criterion is achieved. Following Böhning et al. (1994), the Aitken acceleration criterion has been employed as stopping rule. Given an arbitrary small tolerance value ϵ, this method replaces the standard criterion based on the absolute/relative increment of the log-likelihood with the following stopping rule

$$|l_\infty^{(l+1)} - l_\infty^{(l)}| < \epsilon, \tag{2.3.1}$$

where

$$l_\infty^{(l+1)} = l^{(l)} + \frac{l^{(l+1)} - l^{(l)}}{1 - a^{(l)}} \qquad \text{and} \qquad a^{(l)} = \frac{l^{(l+1)} - l^{(l)}}{l^{(l)} - l^{(l-1)}}$$

indicate respectively the current Aitken accelerated estimate of the maximized log-likelihood and the ratio of two consecutive increments of the log-likelihood. Although the improvement of the log-likelihood in two successive iterations is widely employed in practice, Böhning et al. (1994) consider it a "lack of progress" measure rather than a proper convergence criterion. Such method, for example, could wrongly stop the algorithm in correspondence of a step of the objective function (McNicholas et al., 2010). Criterion (2.3.1), instead, assesses convergence relying on the projection of the optimized log-likelihood, which represents the actual goal of the iterative procedure. For a discussion on the relative merits of the Aitken acceleration criterion and other related proposals see McNicholas et al. (2010).

As far as convergence of the proposed algorithm is concerned, it is difficult to provide theoretical support. In fact, we cannot exploit well-known sufficient conditions provided in Vaida (2005) and McLachlan & Krishnan (2007) due to the discreteness of the ρ parameter component. To our knowledge in the presence of mixed-type parameter space, no previous positive result has been developed. However, we have always experienced a strictly monotone likelihood updating in our algorithm, with stopping rule achieved in a suitable number of iterations. A similarly convincing behavior of the EM algorithm in the presence of mixed-type parameter space is found also in the successful implementations for mixtures of distance-based models in Murphy & Martin (2003) and Lee & Yu (2010).

Another crucial issue in a mixture model setting is the choice of the number of components. In the statistical literature this problem is addressed with several criteria; we opted for the popular *Bayesian Information Criterion*

$$\text{BIC} = -2l(\hat{\theta}_{ML}) + \nu \log N,$$

where $l(\hat{\theta}_{ML})$ is the maximized log-likelihood and ν is the number of free parameters. The BIC, introduced by Schwarz (1978), is a measure which balances between two conflicting goals typically aimed at when fitting a statistical model: good fit and parameter parsimony, where the latter is modulated through the penalty term. In the presence of competing mixture models, the one associated with the lowest BIC value is preferred.

Chapter 3

Ranking models for the Large Fragment Phage Display data

3.1 The LFPD data set

Our investigation is motivated by a real data set coming from a new technology for epitope mapping of the binding between the antibodies present in a biological tissue and a target protein. The biological foundation of the experiment is detailed in Gabrielli et al. (2013) and consists of repeated binding measurements of human blood exposed to $K = 11$ partially overlapping fragments of the HER2 oncoprotein, denoted sequentially by Hum 1,..., Hum 11 (see Figure 3.1). Researchers were originally interested in testing the validity of their innovative biotechnology which consists in a new way of isolating protein fragments without losing the conformational structure of the protein portions. To achieve this goal they employed a phage as a vector for hosting each protein fragment. Then they compared the binding outcome detected on each of the 11 fragments via a standard Enzyme-Linked ImmunoSorbent Assay (ELISA) with the whole protein (Hum 12) and the empty vector (Hum 13) used, respectively, as positive and negative controls (see Figure 3.1). They first checked with monoclonal antibodies that the expected binding at some specific fragment was actually detected. Then they gathered $N = 67$ samples of human blood taken from three different disease groups: i) HD = healthy patients, ii) EBC = patients diagnosed with breast cancer at an early stage, iii) MBC = patients diagnosed with metastatic breast cancer. Binding outcomes from the ELISA experiment have been detected by a laser scanner so that the binding intensities have been measured and recorded in terms of absorbance levels in nanometers (nm). In the next section we motivate our statistical analysis of the LFPD data based on the ordinal information rather than on the original quantitative scale measurements.

3.2 Ranked data modeling of the LFPD data

The original raw absorbance data derived from the LFPD experiment were somehow wildly fluctuating and looked indeed very heterogeneous as apparent in Figure 3.2. However, there were certainly some manifest peaks corresponding to recurrent fragments, especially high for some patients, most frequently those diagnosed

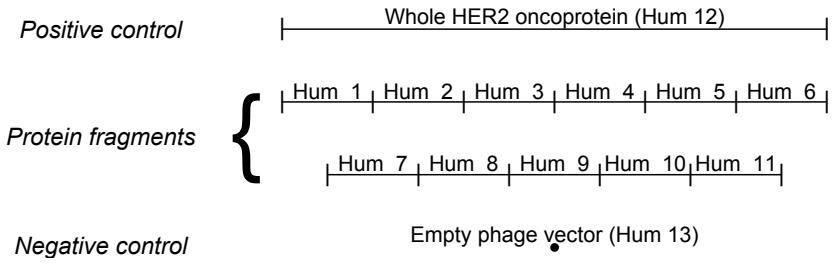

Figure 3.1. 1-D scheme of the HER2 oncoprotein and its segmentation into the 11 partially overlapping fragments (Hum) employed in the LFPD bioassay experiment. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

with cancer. It is also apparent that the individual absorbance profiles are measured at different mean levels for different patients and with different variability. A simple logarithmic transformation and recentering w.r.t. the individual mean were performed providing some more stable evidence of the differential profiles among groups. However, there are some specific profiles which seem pretty much overlapped among different subgroups, although with some different overall pattern (Figure 3.3).

Since data emerged from the development of an innovative technology, miscalibrations or inaccuracies of the measuring device may occur and/or subject-specific characteristics may alter somehow the observed numerical outcome. This makes it more difficult to adjust the statistical analysis based on raw or ad-hoc pre-processed data. Unfortunately, for this kind of data a consolidated and fully-shared normalization technique is lacking. For all these reasons, rather than basing our analysis on the quantitative output of the LFPD study, we verified the possible usefulness of the ranking profiles as a more robust and unambiguously-defined evidence, capable to capture and characterize the sample heterogeneity. Hence, we first derived ordered sequences ranking the absorbance levels of the individual protein fragments taken in decreasing order (rank 1=highest value, rank K=lowest value). We performed a simple exploratory analysis by cancer state computing both the $K \times K$ first-order marginal matrices $\hat{M}$ and the Borda orderings $\overline{\overline{\pi}}^{-1}$. The generic entry $\hat{M}_{ij}$ in the marginal matrix denotes the observed relative frequency that item i is ranked j-th, whereas the sequence $\overline{\overline{\pi}}^{-1}$ lists items taken in order from the highest to the lowest mean rank. These matrices are displayed as image plots in Figure 3.4, together with the Borda sequences at the bottom of each panel. The color intensity is an increasing function of the corresponding observed frequency. The analysis of the first-order marginal matrices suggests that some protein fragments are very often associated with lower ranks, as pointed out by the presence of darker rectangles in correspondence of bottom positions. This constantly occurs for all disease subgroups with Hum 10 but some interesting differential evidence is apparent for EBC subjects with Hum 5 and 6, for MBC with Hum 2 and 13 and also for HD patients with Hum 9 (Figure 3.4). Such a precious discriminant information could be better captured by our EPL. To validate this claim we fitted both the PL and the new EPL to the three disease subgroups separately. For the former we used two known orders, forward (PL-ρ_1) and backward (PL-ρ_2), whereas for the latter

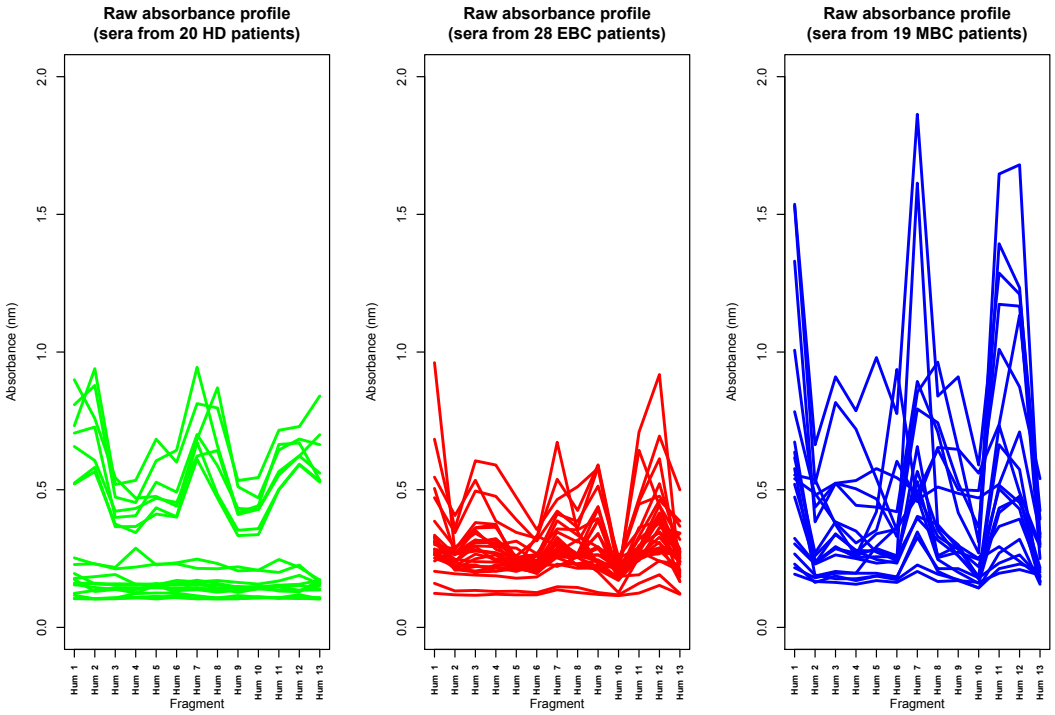

Figure 3.2. Raw absorbance profiles for the three groups of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each continuous piecewise linear function represents the absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

the reference order is a parameter to be estimated. Estimation performances are shown in terms of BIC values in Table 3.1 and reveal that the EPL fit is better or at most comparable with those relative to the PL with fixed reference orders. The interest in relaxing the traditional forward assumption is supported also by the BIC values for the PL-ρ_2, showing that such a model constantly outperforms the PL-ρ_1 when fitted to HD and MBC subjects. These BIC results represent a strong evidence motivating the need of a PL extension. In what follows we consider a more comprehensive analysis in a mixture model setting. With this approach we aim at:

- addressing the heterogeneous nature of the LFPD data using the evidence provided by the orderings of the absorbance levels;

- assessing if and how the path in the sequential ranking process can impact the final model-based classification and select the most appropriate one;

- identifying possible characteristic subgroups related to the disease state;

- characterizing each subgroup with the estimates of the cluster-specific parameters.

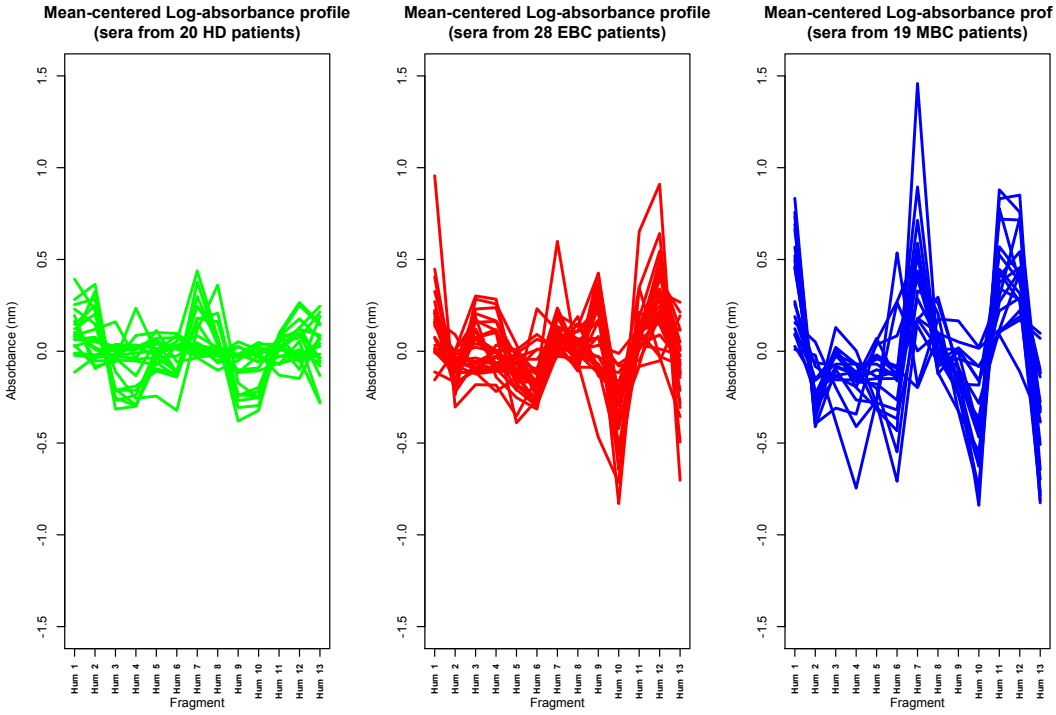

Figure 3.3. Mean-centered log-absorbance profiles for the three groups of patients in the LFPD study: HD = healthy (green), EBC = diagnosed with breast cancer in an early stage (red), MBC = diagnosed with metastatic breast cancer (blue). Each continuous piecewise linear function represents the mean-centered log-absorbance levels in the HER2 oncoprotein fragments (Hum) of a single experimental unit. Hum 12 and Hum 13 indicate respectively the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

3.3 Empirical findings

Considering all the 67 available orderings, we fitted mixtures of DB with $d = d_K$ (DBmix), mixtures of PL with both forward and backward reference order (PLmix-ρ_1 and PLmix-ρ_2), mixtures of BM (BMmix) with dampening parameters shared by all groups as in Gormley & Murphy (2008a) and mixtures of EPL. In the most general mixture model (EPLmix) each EPL component has a group-specific parameter ρ_g to be inferred. We have also considered a constrained version with an unknown reference order ρ common to all components and we have denoted it with PLmix-ρ. All mixtures have been implemented with a number of components varying from $G = 1$ to $G = 7$. Of course, the case $G = 1$ coincides with the assumption that observations come from a homogeneous population without an underlying group structure. We separately applied all mixture models to the ranked absorbance levels relative to the $K = 11$ partially overlapping protein fragments as well as to the $K = 11 + 2$ binding probes (spots), including also the whole HER2 oncoprotein (positive control) and the empty phage vector (negative control).

Focusing on the BIC for $G = 1$ compared to $G > 1$, the MLE of the DBmix provided an overall evidence in favor of heterogeneity when both $K = 11$ or $K = 13$ binding probes are considered. We highlighted a remarkably decreasing behavior for

Table 3.1. BIC values and corresponding differences ΔBIC w.r.t. the best fitting model. MLE of PL-ρ_1, PL-ρ_2 and EPL have been computed separately for the three disease groups (HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer) and for a different number K of binding probes included in the rankings. The smallest BIC and ΔBIC values indicating the best fitted models are highlighted in bold.

Model	$K = 11$					
	HD		EBC		MBC	
	BIC	ΔBIC	BIC	ΔBIC	BIC	ΔBIC
PL-ρ_1	694.04	17.11	776.13	2.96	499.46	25.56
PL-ρ_2	685.85	8.92	804.61	31.44	498.67	24.77
EPL	**676.93**	**0**	**773.17**	**0**	**473.90**	**0**
	$K = 13$					
	HD		EBC		MBC	
	BIC	ΔBIC	BIC	ΔBIC	BIC	ΔBIC
PL-ρ_1	899.63	25.92	**1025.71**	**0**	658.44	28.39
PL-ρ_2	894.44	20.73	1039.45	13.74	652.15	22.10
EPL	**873.71**	**0**	1026.61	0.90	**630.05**	**0**

Table 3.2. BIC values resulting from the MLE of the DBmix on the LFPD data with a varying number G of components, when either $K = 11$ or $K = 13$ binding probes are included in the rankings.

	G									
	1	2	3	4	5	6	7	8	9	10
$K = 11$	2078.77	2003.65	1940.86	1899.33	1882.32	1863.17	1846.98	1829.81	1817.06	1798.12
$K = 13$	2700.02	2617.66	2551.38	2512.19	2483.25	2451.38	2421.71	2392.10	2366.78	2342.60

the associated BIC, which persists even when the fitting is carried out up to $G = 10$ components as shown in Table 3.2. Indeed, fitting DBmix with an increasing number of groups pointed out a particular feature of the DB, probably due to the sparse nature of LFPD data. We remind, in fact, that in the present application the sample size is small w.r.t. to the cardinality (11! or 13!) of the discrete ranking space. As the value of G in the DBmix increases, some components start to represent just a single observation. This can be explained, perhaps, by the fact that, once the modal ranking σ has been fixed, the DB has only one parameter left for fitting the amount of uncertainty around σ. It follows that for these components the concentration parameter λ is typically estimated as a very high value. This behavior, of course, could make the DBmix model not sufficiently parsimonious and suitable in some sparse-data situations because it could lead to a more sparse clustering of the observations and to a less enlightening inferential findings.

When stagewise models were fitted to the LFPD data, we found again evidence in favor of the heterogeneous structure. Since in the comparison between the Kendall and the Cayley distance employed in the local search step neither of the two metrics yielded a consistently better solution, we decided to report in Table 3.3 only the BIC

Table 3.3. BIC values, corresponding differences ΔBIC w.r.t. the best fitting model and number G of components of the best PLmix-ρ_1, PLmix-ρ_2, BMmix, PLmix-ρ and EPLmix fitted to LFPD data for different number K of binding probes included in the ranking. The smallest BIC and ΔBIC values indicating the best fitted models are highlighted in bold.

	$K = 11$			$K = 13$		
Model	BIC	ΔBIC	G	BIC	ΔBIC	G
PLmix-ρ_1	1964.87	29.10	4	2589.93	71.37	5
PLmix-ρ_2	1984.53	48.76	4	2619.60	101.04	4
BMmix	1995.09	59.32	4	2615.45	96.89	5
PLmix-ρ	1969.24	33.47	4	2597.42	78.86	5
EPLmix	**1935.77**	**0**	4	**2518.56**	**0**	5

values obtained with the Kendall distance. In the case $K = 11$ all types of mixtures consistently identify four groups in the sample. When also the control probes are included in the ordered sequences, five groups are consistently selected with the only exception of PLmix-ρ_2. Bold BIC values in Table 3.3 point out the EPLmix as the best model. Optimal BIC values of the EPLmix are, in both cases, significantly smaller than those corresponding to the competing mixtures, as also stressed by ΔBIC values in the same table. Indeed, the outperformance of EPLmix for any fixed G of the mixture is apparent in Figure 3.5. Hence, the introduction of the discrete parameter in the mixture component leads to a remarkable improvement of fit when it is allowed to be variable among groups. Note, in fact, the gap between the BIC trend associated to the more flexible EPLmix and the one relative to its restricted version PLmix-ρ. Comparing also the EPLmix with the BMmix, BIC results apparently show that the larger flexibility provided by the BM does not lead to an improvement, since with these data the penalization for a larger number of parameters exceeds the gain in terms of log-likelihood.

The selected EPLmix exhibits a good accuracy in the discrimination of sample units w.r.t. the real disease state. The two resulting clusterings, in fact, agree with the most relevant distinction of the real disease state between healthy and unhealthy patients, as pointed out in Tables 3.4(a) and 3.4(b). Specifically, collapsing the model-based group membership into the above basic bipartition, we recognize that healthy subjects are well isolated, with only 1 or 2 false positive cases; for diseased patients, instead, we have 7 misclassifications in the $K = 11$ case but only 2 with the addition of the control spots, see Tables 3.4(a) and 3.4(b). As expected, the inclusion of the positive and negative controls provided an additional discriminating power, measured by the increment of the Adjusted Rand Index (ARI) from 0.52 to 0.77. Healthy patients are always modeled by two components in all the fitted mixtures. This hints at possibly different subtypes of healthy profiles. In fact, we can easily verify that such subdivision reflects two different absorbance patterns in cancer-free units, made evident in Figure 3.2 by the green continuous piecewise linear functions: a first subgroup whose immune defenses essentially did not react at all to the exposition to the HER2 oncoprotein (lower panel) and a second one with

Table 3.4. Correspondence between the model-based clustering derived from the MLE of the EPLmix and the true disease state of the LFPD experimental units: HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer. Corresponding values for the Adjusted Rand Index (ARI) based on the basic healthy/unhealthy bipartition are shown in parentheses.

(a) $K = 11 \quad (ARI = 0.52)$

	Group			
Disease state	1	2	3	4
HD	0	2	10	8
EBC	13	12	2	1
MBC	0	15	3	1

(b) $K = 13 \quad (ARI = 0.77)$

	Group				
Disease state	1	2	3	4	5
HD	1	10	0	1	8
EBC	12	0	9	7	0
MBC	14	2	0	3	0

some manifest and characterized binding profile (upper panel). On the other hand, among the components representing diseased patients, the sub-classification between EBC and MBC is only partially recovered, especially for the latter subgroup. This is proved by the presence of at least one model-based group entirely composed of EBC subjects in all the fitted mixtures, whereas MBC patients always belong to mixed-type components.

The varying correspondence between the real cancer state and the inferred clustering structure confirms the presumed dependence of the classification results on the adopted reference ranking process ρ. Furthermore, the good agreement obtained with the EPLmix (Tables 3.4(a) and 3.4(b)) suggests that researchers should not focus exclusively on differential epitope identification but could extend their analysis considering also a more general global understanding of differential bindings. Hence, in order to characterize disease groups w.r.t. ranking profiles, it is interesting to interpret the component-specific modal orderings (Table 3.5), derived by ordering the corresponding support parameter estimates (Figure 3.6). Results refer to PLmix-ρ_1, PLmix-ρ_2 and EPLmix; inferential findings for BMmix and EPL-ρ were very similar to PLmix-ρ_1 and are not shown. Weights and reference order estimates of the identified clusters are shown in Table 3.6. Focusing on the analysis based on 13 binding probes, we stress that in the best fitted models the positive control probe (Hum 12) repeatedly occupies top positions in the modal orderings of EBC and EBC+MBC mixture components. We remind that Hum 12 denotes the absorbance level corresponding to the entire HER2 oncoprotein. Thus, in theory, its level should reflect the total binding and it is reasonably expected to be higher than absorbance level detected in limited portions of the oncoprotein. On the other hand, immunological response in healthy patients may either be unaffected by the exposition to the HER2 oncoprotein or yield a mild binding. This implies an exchangeability of binding probes in the ordering of absorbance levels which is typical of the UM. These aspects reinforce the presence of Hum 12 in top positions as a signal that the immunological response actually occurred and it can be reasonably interpreted as a distinguishing feature of the unhealthy patients. It turns out that with our wildly fluctuating LFPD data it would not be possible to identify a simple threshold for the raw (or normalized) binding outcome to discriminate unhealthy patients. This objective is better achieved using binding profiles based on rankings.

Table 3.5. Modal orderings derived from the best PLmix-ρ_1, PLmix-ρ_2 and EPLmix fitted to LFPD data for a different number K of binding probes included in the rankings. "D.C." stands for "disease composition" and lists sequentially the number of HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer patients in each group. The symbol $*$ indicates mixture components which are very close to the UM.

	$K=11$			$K=13$		
Model	g	D.C.	$\hat{\sigma}_g^{-1}$	g	D.C.	$\hat{\sigma}_g^{-1}$
PLmix-ρ_1	1	$(11,1,3)$	$(6,1,5,4,7,11,3,8,10,9,2)$*	1	$(2,11,3)$	$(12,1,11,7,8,9,2,13,3,4,6,5,10)$
	2	$(0,10,0)$	$(9,3,7,11,4,1,8,2,5,6,10)$	2	$(6,0,0)$	$(7,2,1,12,8,11,13,5,6,10,4,3,9)$
	3	$(3,17,15)$	$(1,11,7,8,3,9,4,5,2,6,10)$	3	$(0,7,0)$	$(9,3,4,12,11,7,1,8,13,2,5,6,10)$
	4	$(6,0,1)$	$(7,2,1,8,11,5,6,4,10,3,9)$	4	$(0,7,14)$	$(1,12,11,7,8,3,5,9,4,6,2,13,10)$
				5	$(12,3,2)$	$(1,5,6,7,3,4,11,12,8,9,10,2,13)$*
PLmix-ρ_2	1	$(14,5,3)$	$(1,6,11,7,5,8,2,4,9,3,10)$*	1	$(12,2,2)$	$(1,6,12,5,7,11,4,2,13,3,10,9,8)$*
	2	$(6,0,1)$	$(7,2,1,8,11,5,6,3,4,10,9)$	2	$(0,15,0)$	$(12,9,3,7,4,11,1,13,8,2,5,6,10)$
	3	$(0,12,15)$	$(1,7,11,8,3,9,4,5,2,6,10)$	3	$(1,11,17)$	$(12,1,11,7,8,3,9,5,6,4,13,2,10)$
	4	$(0,11,0)$	$(9,3,4,7,11,1,8,2,5,6,10)$	4	$(7,0,0)$	$(7,2,1,12,8,13,11,5,6,3,4,10,9)$
EPLmix	1	$(0,13,0)$	$(9,8,1,3,11,7,2,4,5,6,10)$	1	$(1,12,14)$	$(12,1,11,7,8,3,4,9,5,6,2,13,10)$
	2	$(2,12,15)$	$(1,11,7,8,9,3,4,5,6,2,10)$	2	$(10,0,2)$	$(5,2,11,4,3,6,10,7,8,9,12,1,13)$
	3	$(10,2,3)$	$(5,4,11,1,6,3,10,2,9,7,8)$	3	$(0,9,0)$	$(9,12,11,3,1,4,7,2,13,8,5,6,10)$
	4	$(8,1,1)$	$(7,2,1,8,11,5,6,4,10,9,3)$	4	$(1,7,3)$	$(12,9,1,11,13,3,8,7,2,4,5,6,10)$
				5	$(8,0,0)$	$(11,2,1,6,12,8,13,5,7,10,4,3,9)$

Moreover, the combination of Hum 12 with the pattern (Hum 1, Hum 11, Hum 7) in top positions seems to characterize mixed (EBC+MBC) diseased groups, such as the first and the fourth components in PLmix-ρ_1, the third one in PLmix-ρ_2 and the first one in EPLmix. In fact, the protein fragments Hum 1, Hum 11 and 7 were already recognized in Gabrielli et al. (2013) as the relevant epitopes. Referring to EBC-specific components, similar results are valid for the fragment pair (Hum 9, Hum 3) which, together with the positive control, occupies the very first top positions (see for example the third group in PLmix-ρ_1 and the second one in PLmix-ρ_2). This means that for some EBC patients the binding reaction mainly occurs in a different section of the oncoprotein, improving the discrimination of this subgroup among diseased patients. Relevant findings can be also highlighted for healthy patients. The absent or negligible immunological response observed for some of them is well described in the estimated models by the presence of a component which is very close to the UM, as shown by the corresponding inferred values $\underline{\hat{p}}_g$. In this case the modal orderings are poorly representative, so we marked them with the symbol $*$ in Table 3.5. These UM-like components involve prevalently HD patients. They are also involved in another more characterized mixture component. The interpretation of the non-uniform component parameters suggests that some HD subjects share the epitopes Hum 1 and Hum 7 with other patients but they also have a distinctive Hum 2 in top positions; Hum 11, instead, appears in middle positions. We can also look at low absorbance patterns, if bottom ranks can be regarded as meaningful *signatures* for the problem at hand. Note, for example, that while Hum 10 appears consistently in last positions for almost all of the fitted components, Hum 9 seems to be a sort of "anti"-epitope signature for HD units. The

Table 3.6. Mixture weights and reference order estimates of the best PLmix-ρ_1, PLmix-ρ_2 and EPLmix fitted to LFPD data for a different number K of binding probes included in the rankings.

	$K = 11$			$K = 13$	
Model	g	$\hat{\omega}_g$	$\hat{\rho}_g$	g	$\hat{\omega}_g$	$\hat{\rho}_g$
PLmix-ρ_1	1	.22	ρ_1	1	.24	ρ_1
	2	.15	ρ_1	2	.09	ρ_1
	3	.53	ρ_1	3	.11	ρ_1
	4	.10	ρ_1	4	.31	ρ_1
				5	.25	ρ_1
PLmix-ρ_2	1	.35	ρ_2	1	.25	ρ_2
	2	.10	ρ_2	2	.22	ρ_2
	3	.39	ρ_2	3	.43	ρ_2
	4	.16	ρ_2	4	.10	ρ_2
EPLmix	1	.19	$(11, 10, 9, 7, 8, 4, 2, 3, 6, 5, 1)$	1	.39	$(2, 1, 3, 4, 5, 6, 8, 9, 7, 10, 11, 12, 13)$
	2	.44	$(1, 2, 3, 4, 6, 5, 7, 8, 9, 10, 11)$	2	.18	$(6, 9, 2, 12, 13, 4, 8, 1, 3, 7, 11, 5, 10)$
	3	.22	$(6, 9, 7, 10, 4, 5, 8, 2, 1, 11, 3)$	3	.14	$(12, 11, 8, 10, 9, 5, 7, 6, 3, 4, 2, 1, 13)$
	4	.15	$(3, 1, 2, 4, 5, 9, 11, 10, 8, 7, 6)$	4	.17	$(1, 4, 3, 7, 8, 2, 9, 5, 6, 10, 12, 13, 11)$
				5	.12	$(8, 13, 12, 10, 11, 1, 6, 7, 4, 5, 9, 2, 3)$

same role is played by the Hum pattern (Hum 5, Hum 6) for EBC subjects. Another interesting feature regards Hum 13; it corresponds to the empty phage vector and hence, theoretically, one would expect it to be associated with bottom ranks. This is true, instead, for those groups composed for the most part of MBC units (see for example the fourth component in the PLmix-ρ_1, the third one in the PLmix-ρ_2 and the first one in the EPLmix). Therefore, a minimum absorbance level in Hum 13 could be an important feature to discriminate MBC patients, the subgroup which is only weakly characterized by the present analysis. Similar observations are valid for the case $K = 11$ omitting, naturally, Hum 12 and Hum 13. Finally we remark that the EPLmix, selected as the best model in terms of the BIC, involves 69 parameters in the case of $K = 13$.

3.4 Alternative quantitative data analysis

In this section we show that our analysis based on ranked data and the EPL mixture model compares favorably with a more conventional quantitative data approach. We implemented the flexible mixture of multivariate normal distributions (MNorm-mix) with the R package `mclust` described in Fraley & Raftery (2003).

As urged in Section 3.2, we must preliminarily decide whether there exists a more appropriate way of transforming and rescaling the original quantitative measures. Since a consolidated normalization method is lacking for this type of experiments, we worked with 3 alternative reasonable options: original raw data, the log-transformed absorbances and the rescaled log-transformed absorbances so that

Table 3.7. Correspondence between the model-based clustering derived from the MLE of the MNorm-mix and the true disease state of the LFPD experimental units: HD = healthy, EBC = diagnosed with early stage breast cancer and MBC = diagnosed with metastatic breast cancer. Corresponding values for the Adjusted Rand Index (ARI) based on the basic healthy/unhealthy bipartition are shown in parentheses.

(a) Raw LFPD data with 11 Hum ($ARI = 0.32$)

	Group		
Disease state	1	2	3
HD	7	9	4
EBC	3	2	23
MBC	9	0	10

(b) Rescaled log-transformed LFPD data with 13 Hum ($ARI = 0.88$)

	Group						
Disease state	1	2	3	4	5	6	7
HD	7	11	2	0	0	0	0
EBC	0	0	12	1	10	5	0
MBC	0	0	3	5	0	9	2

the individual average log-absorbance of all the considered spots is null for each patient. Results derived from the quantitative analysis are very different according to the measurement scale adopted in the input data. In fact, only with the raw data the best fitting mixture model provides evidence in favor of an heterogeneous model, namely a mixture with $G = 3$ components. However, as shown in Table 3.7(a), the correspondence with the known disease state is poorer ($ARI = 0.32$) than the one obtained with the ranking-based analysis. In all other cases the MNorm-mix model selected the single component homogeneous model as the best fitting. However, if the model is forced to be fitted as heterogeneous, then a variable number of groups is selected, ranging from 4 to 7. Indeed, the best classification is the one obtained with a MNorm-mix applied to the rescaled log-transformed absorbances of all the 13 fragments. This grouping has a very good agreement with the three disease subgroups ($ARI = 0.88$), as shown in Table 3.7(b). However, we stress that this model does not represent the best fitting in terms of BIC and yields a more scattered clustering. Moreover, this model requires 117 parameters; hence, it is less parsimonious and can be more difficult to interpret than the best fitting mixture for ranked data.

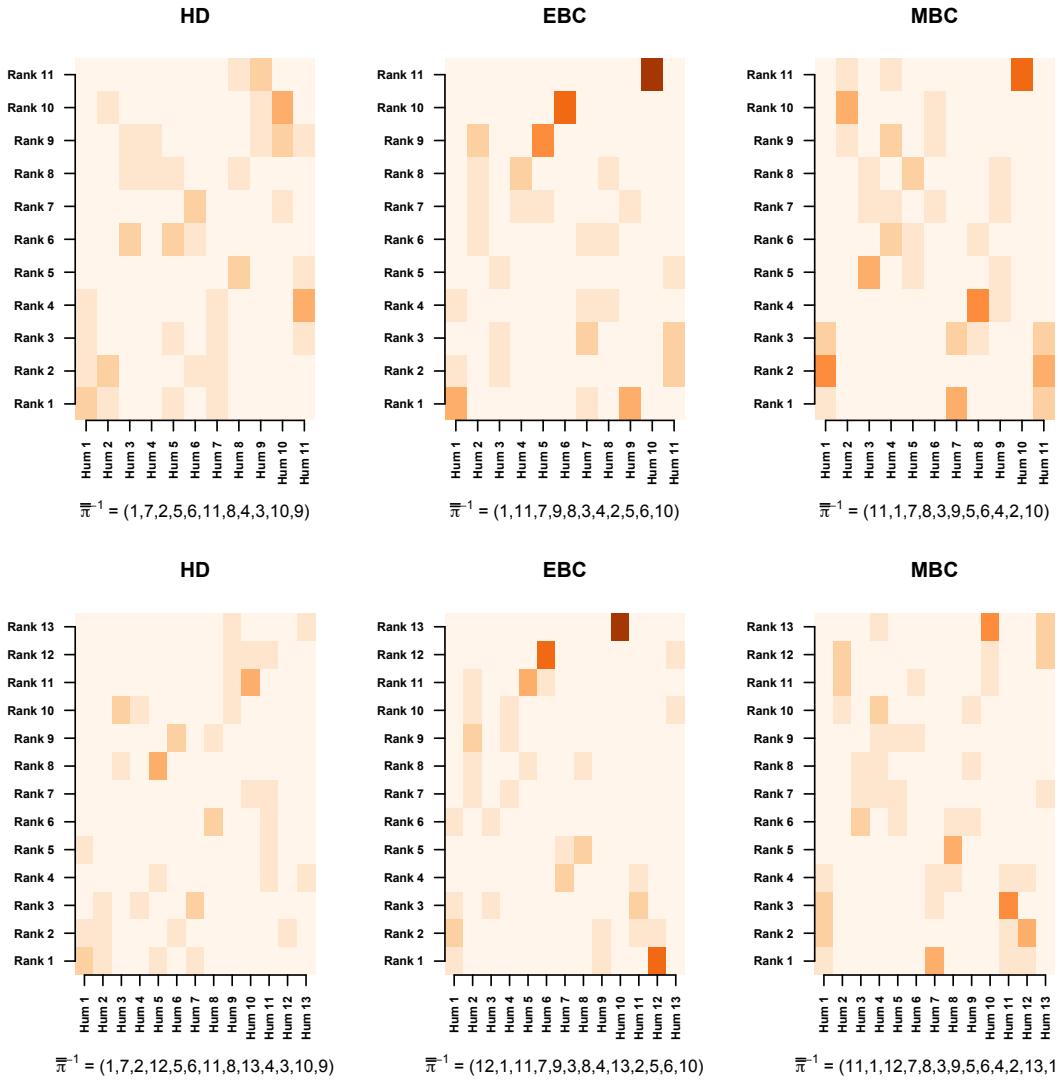

Figure 3.4. Image plots of the first-order marginal matrices for the three groups of patients in the LFPD study: HD = healthy (left), EBC = diagnosed with early stage breast cancer (center), MBC = diagnosed with metastatic breast cancer (right). Upper panel refers to the data with $K = 11$ protein fragments, whereas lower panel concerns the case with $K = 13$ binding probes, including the whole HER2 oncoprotein (Hum 12 = positive control) and the empty phage vector (Hum 13 = negative control). The Borda ordering $\overline{\overline{\pi}}^{-1}$ lists items taken in order from the highest to the lowest mean rank.

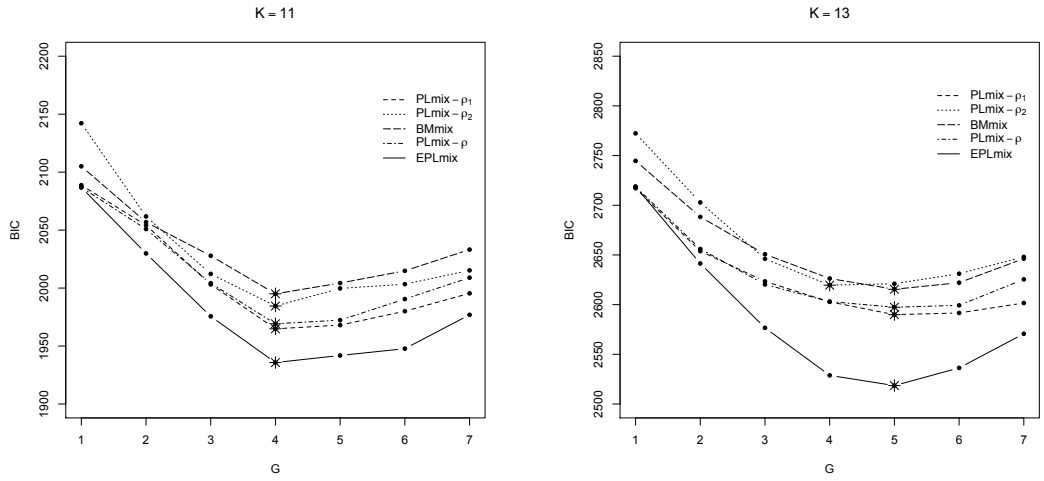

Figure 3.5. BIC trends resulting from the MLE of the PLmix-ρ_1, PLmix-ρ_2, BMmix, PLmix-ρ and EPLmix on the LFPD data with varying number G of mixture components, when either $K = 11$ (left) or $K = 13$ (right) binding probes are included in the ranking. The symbol ✶ indicates the minimum BIC values for the final selection of the number of groups.

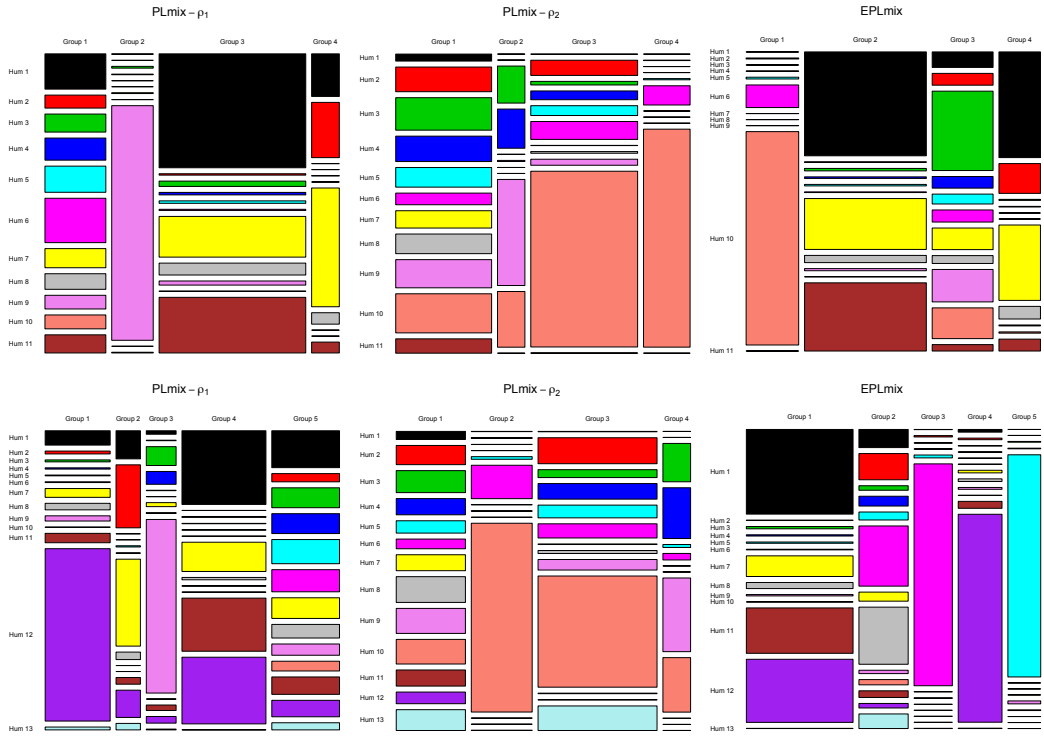

Figure 3.6. Support parameter estimates represented via mosaic plots for the best PLmix-ρ_1, PLmix-ρ_2 and EPLmix fitted to the LFPD data. Bar widths are proportional to group weights. Upper panel refers to the data with $K = 11$ protein fragments, whereas lower panel concerns the case with $K = 13$ binding probes, including the whole HER2 oncoprotein (Hum 12 = positive control) and the empty phage vector (Hum 13 = negative control).

Chapter 4

Benterized Extended Plackett-Luce model for ranked data

In the light of the approving fitting obtained by the implementation of the novel EPL to the LFPD data, we considered the possibility to extend further this parametric class. The methodological contribution developed in this chapter moves from merging the EPL with another well-established PL extension, such as the BM. These models, in fact, describe substantially different but complementary attributes of the ranking procedure. In fact, selection accuracy and reference order in the ranking process could be combined to construct a more flexible parametric PL generalization, which incorporates both the EPL and the BM as specific parameter settings. In the following sections we give the formal definition of the new statistical model, with inferential details for its MLE implementation in both the homogeneous and heterogeneous population approaches.

4.1 Benterized Extended Plackett-Luce model

The straightforward way to involve the uncertainty related to the item selection process in the EPL definition given by (1.3.1) is assuming that the probability of a generic ordering has the following form

$$\mathbf{P}_{\text{BEPL}}(\pi^{-1}|\rho,\underline{p},\underline{\alpha}) = \mathbf{P}_{\text{BENT}}(\pi^{-1}\rho|\underline{p},\underline{\alpha}) = \prod_{t=1}^{K} \frac{p_{\pi^{-1}(\rho(t))}^{\alpha_t}}{\sum_{\nu=t}^{K} p_{\pi^{-1}(\rho(\nu))}^{\alpha_t}} \qquad \pi^{-1} \in \mathcal{S}_K,$$

which we name *Benterized Extended Plackett-Luce model* and indicate in short with BEPL(ρ,$\underline{p}$,$\underline{\alpha}$). Note the different effect of ρ and $\underline{\alpha}$ in the above probabilistic construction: the former determines the order in the sequential normalization of the support parameters, whereas the latter perturb them. The motivating subclasses (EPL, BM and PL) are easily recovered with suitable parameter configurations. Setting $\alpha_t = 1$ for all t, the BEPL becomes the EPL, whereas assuming $\rho = e$ we obtain the BM. Finally considering jointly such constraints on the dampening and the reference order parameters and letting only the support parameters vary, the model reduces to the ordinary PL.

4.2 MLE of the Benterized Extended Plackett-Luce model

We first consider the homogenous population case ($G = 1$). Let BEPL(ρ,$\underline{p}$,$\underline{\alpha}$) be the underlying mechanism generating the observed orderings $\underline{\pi}^{-1}$ and implying the following expression for the log-likelihood

$$l(\rho, \underline{p}, \underline{\alpha}) = \sum_{s=1}^{N} \sum_{t=1}^{K} \left(\alpha_t \log p_{\pi^{-1}(\rho(t))} - \log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t} \right). \tag{4.2.1}$$

The optimization of the log-likelihood is again based on the iterative MM algorithm mimicing the approach adopted in Gormley & Murphy (2008a) to infer the BM. We extend such a procedure with the presence of the reference order parameter and derive the corresponding estimation formula. The estimation of the continuous parameters, $\underline{p}$ and $\underline{\alpha}$, requires a double minorization step to be applied to $-\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}$ and subsequently to $-p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}$, which have to be interpreted as function of $\underline{p}$ and $\underline{\alpha}$ according to which quantity we want to focus on. For the support parameters, the surrogate is obtained by exploiting the convexity of both functions and hence, as seen for the EPL, the supporting hyperplane property. From Taylor's first-order expansion one has

$$-\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t} \geq 1 - \log \sum_{\nu=t}^{K} \overline{p}_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t} - \frac{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}}{\sum_{\nu=t}^{K} \overline{p}_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}},$$

where also the term $-p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}$ is in turn minorized with its linear approximation

$$-p^{\alpha} \geq -\overline{p}^{\alpha} - \alpha\, \overline{p}^{\alpha-1}(p - \overline{p}).$$

Disregarding additive terms not depending on $\underline{p}$ and plugging-in (4.2.1), the final auxiliary objective function turns out to be

$$q = \sum_{s=1}^{N} \sum_{t=1}^{K} \alpha_t \log p_{\pi^{-1}(\rho(t))} - \sum_{s=1}^{N} \sum_{t=1}^{K} \frac{\sum_{\nu=t}^{K} \alpha_t \overline{p}_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t - 1} p_{\pi_s^{-1}(\rho(\nu))}}{\sum_{\nu=t}^{K} \overline{p}_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}}.$$

Differentiating w.r.t. each p_i and equating the partial derivatives to zero, the estimation formula for the support parameters at the current iteration of the MM algorithm can be written as

$$p_i^{(l+1)} = \frac{\sum_{s=1}^{N} \sum_{t=1}^{K} \alpha_t^{(l)} \xi_{sti}^{(l)}}{\sum_{s=1}^{N} \sum_{t=1}^{K} \dfrac{\alpha_t^{(l)} \left(p_i^{(l)}\right)^{\alpha_t^{(l)} - 1} \delta_{sti}^{(l)}}{\sum_{\nu=t}^{K} \left(p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l)} \right)^{\alpha_t^{(l)}}}} \qquad i = 1, \ldots, K,$$

where

$$\xi_{sti}^{(l)} = \begin{cases} 1 & \text{if } i = \pi_s^{-1}(\rho^{(l)}(t)), \\ 0 & \text{otherwise} \end{cases}$$

and $\delta_{sti}^{(l)} = \sum_{\nu=t}^{K} \xi_{s\nu i}^{(l)}$ corresponds to the binary array already defined in Section 2.2. Notice that we have replaced with $p^{(l)}$ the quantity $\overline{p}$ in the previous minorizations. The presence of the additional binary array $\xi^{(l)}$ is justified by the dampening parameters which induce the dependence of the numerator of the likelihood on the stagewise order of selections. Its generic element $\xi_{sti}^{(l)}$ is the indicator of the event that unit s has chosen item i at stage t.

Also the dampening parameter estimates are derived applying a double minorization, where the first one is essentially the same step described for the p's but with α treated as the variable of interest:

$$-\log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t} \geq 1 - \log \sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\overline{\alpha}_t} - \frac{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t}}{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\overline{\alpha}_t}}.$$

In this case the function $-p^{\alpha}$ is concave in α but we can operate in the following way: first we bound the convex p^{α} around $\overline{\alpha}$ with the quadratic function given by

$$p^{\alpha} \leq p^{\overline{\alpha}} + (\alpha - \overline{\alpha})(\log p) p^{\overline{\alpha}} + \frac{(\alpha - \overline{\alpha})^2}{2} (\log p)^2,$$

where $(\log p)^2 \geq (\log p)^2 p^{\overline{\alpha}} = d^2 p^{\overline{\alpha}} / d\overline{\alpha}^2$; then, one simply reverses the inequality

$$-p^{\alpha} \geq -p^{\overline{\alpha}} - (\alpha - \overline{\alpha})(\log p) p^{\overline{\alpha}} - \frac{(\alpha - \overline{\alpha})^2}{2} (\log p)^2.$$

It follows that the formula for the minorizing surrogate function is

$$\begin{aligned} q = & \sum_{s=1}^{N} \sum_{t=1}^{K} \alpha_t \log p_{\pi_s^{-1}(\rho(t))} \\ & - \sum_{s=1}^{N} \sum_{t=1}^{K} \frac{1}{\sum_{\nu=t}^{K} p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t^{(l)}}} \sum_{\nu=t}^{K} \Bigg((\alpha_t - \alpha_t^{(l)}) p_{\pi_s^{-1}(\rho(\nu))}^{\alpha_t^{(l)}} \log p_{\pi_s^{-1}(\rho(\nu))} \\ & + \frac{(\alpha_t - \alpha_t^{(l)})^2}{2} (\log p_{\pi_s^{-1}(\rho(\nu))})^2 \Bigg). \end{aligned}$$

Equating the partial derivatives w.r.t. each α_t to zero, the updating rule at the current iteration for the dampening parameters is

$$\alpha_t^{(l+1)} = \alpha_t^{(l)} + \frac{\sum_{s=1}^{N} \frac{\sum_{\nu=t}^{K} \left(p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l+1)}\right)^{\alpha_t^{(l)}} \left(\log p_{\pi_s^{-1}(\rho^{(l)}(t))}^{(l+1)} - \log p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l+1)}\right)}{\sum_{\nu=t}^{K} \left(p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l+1)}\right)^{\alpha_t^{(l)}}}}{\sum_{s=1}^{N} \frac{\sum_{\nu=t}^{K} \left(\log p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l+1)}\right)^2}{\sum_{\nu=t}^{K} \left(p_{\pi_s^{-1}(\rho^{(l)}(\nu))}^{(l+1)}\right)^{\alpha_t^{(l)}}}}$$

for $t = 2, \ldots, K-1$. Using the original log-likelihood, the current estimate of the reference order parameter is derived as

$$\rho^{(l+1)} = \arg\max_{\rho} \, l(\rho, \underline{p}^{(l+1)}, \underline{\alpha}^{(l+1)}).$$

4.3 MLE of the mixture of Benterized Extended Plackett-Luce models

Now we contemplate a more flexible mixture model with BEPL components and discuss the related inferential issues. Augmenting data with the missing group membership indicators $\underline{z}$, one obtains the following expression for the complete log-likelihood

$$l_c(\underline{\rho}, \underline{p}, \underline{\omega}, \underline{z}) = \sum_{s=1}^{N}\sum_{g=1}^{G} z_{sg}\Bigg(\log\omega_g + \sum_{t=1}^{K}\alpha_t \log p_{g\pi^{-1}(\rho_g(t))} - \sum_{t=1}^{K}\log\sum_{\nu=t}^{K} p^{\alpha_t}_{g\pi_s^{-1}(\rho_g(\nu))}\Bigg).$$

Note that in the present model setup the dampening parameters $\underline{\alpha}$ are assumed to be constant among clusters, as in Gormley & Murphy (2008a) and Gormley & Murphy (2008b). In analogy with the MLE of the EPL mixture described in Section 2.2, the maximization of the likelihood is achieved with the hybrid EMM procedure. The E-step leads to

$$\begin{aligned} Q((\underline{\rho}, \underline{p}, \underline{\omega}), (\underline{\rho}^*, \underline{p}^*, \underline{\omega}^*)) &= \mathbb{E}[l_c|\underline{\pi}^{-1}, \underline{\rho}^*, \underline{p}^*, \underline{\omega}^*] \\ &= \sum_{s=1}^{N}\sum_{g=1}^{G} \hat{z}_{sg}\Bigg(\log\omega_g + \sum_{t=1}^{K}\alpha_t \log p_{g\pi^{-1}(\rho_g(t))} \\ &\quad - \sum_{t=1}^{K}\log\sum_{\nu=t}^{K} p^{\alpha_t}_{g\pi_s^{-1}(\rho_g(\nu))}\Bigg), \end{aligned}$$

where

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)}\mathbf{P}_{\mathrm{BEPL}}(\pi_s^{-1}|\rho_g^{(l)}, \underline{p}_g^{(l)}, \underline{\alpha}^{(l)})}{\sum_{g'=1}^{G}\omega_{g'}^{(l)}\mathbf{P}_{\mathrm{BEPL}}(\pi_s^{-1}|\rho_{g'}^{(l)}, \underline{p}_{g'}^{(l)}, \underline{\alpha}^{(l)})}$$

for $s = 1, \ldots, N$ and $g = 1, \ldots, G$. The M-step for the support and dampening parameters requires the derivation of specific surrogate objective functions according to the considerations already made for the homogeneous population case. Thus, in order to avoid redundant details, we only provide the final expressions for the parameter updates. For further details see Gormley & Murphy (2008a). The updating of p_{gi} at the current iteration is

$$p_{gi}^{(l+1)} = \frac{\sum_{s=1}^{N}\hat{z}_{sg}\sum_{t=1}^{K}\alpha_t^{(l)}\xi_{stig}^{(l)}}{\sum_{s=1}^{N}\hat{z}_{sg}\sum_{t=1}^{K}\dfrac{\alpha_t^{(l)}(p_{gi}^{(l)})^{\alpha_t^{(l)}-1}\delta_{stig}^{(l)}}{\sum_{\nu=t}^{K}\left(p^{(l)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)^{\alpha_t^{(l)}}}}$$

for $g = 1, \ldots, G$ and $i = 1, \ldots, K$, with

$$\xi_{stig}^{(l)} = \begin{cases} 1 & \text{if } i = \pi_s^{-1}(\rho_g^{(l)}(t)), \\ 0 & \text{otherwise} \end{cases}$$

and $\delta_{stig}^{(l)} = \sum_{\nu=t}^{K} \xi_{s\nu ig}^{(l)}$ as in Section 2.2. The estimation formula for the dampening parameters shared by the clusters turns out to be

$$\alpha_t^{(l+1)} = \alpha_t^{(l)} + \frac{\sum_{s=1}^{N}\sum_{g=1}^{G} \hat{z}_{sg} \dfrac{\sum_{\nu=t}^{K} \left(p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)^{\alpha_t^{(l)}} \left(\log p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(t))} - \log p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)}{\sum_{\nu=t}^{K} \left(p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)^{\alpha_t^{(l)}}}}{\sum_{s=1}^{N}\sum_{g=1}^{G} \hat{z}_{sg} \dfrac{\sum_{\nu=t}^{K} \left(\log p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)^2}{\sum_{\nu=t}^{K} \left(p^{(l+1)}_{g\pi_s^{-1}(\rho_g^{(l)}(\nu))}\right)^{\alpha_t^{(l)}}}}$$

for $t = 2, \ldots, K-1$, whereas for the reference orders in each subgroup one has to find the solution of the optimization problems given, for $g = 1, \ldots, G$, by

$$\rho_g^{(l+1)} = \arg\max_{\rho_g} \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} \log \mathbf{P}_{\mathrm{BEPL}}(\pi_s^{-1} | \rho_g, \underline{p}_g^{(l+1)}, \underline{\alpha}^{(l+1)}),$$

which can be possibly solved with a local search. The M-step ends with the computation of the mixture weights estimated as

$$\omega_g^{(l+1)} = \frac{\sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}{N} \qquad g = 1, \ldots, G.$$

Chapter 5

Bayesian inference for ranking models

Within the Bayesian paradigm the unknown parameters of interest are treated as random entities and the overall information on them produced by the experiment is expressed in stochastic form through the *posterior* distribution. This combines the prior belief on the parameters, formalized through the *prior* distribution, and the evidence from the data represented by the *likelihood* function, which plays also a central role in the frequentist approach. In this framework inference is conducted employing the final (posterior) probability distribution and suitable summaries thereof.

For a large variety of models the exact computation of the posterior is unfeasible. This fact explains the wide assortment of techniques proposed in the Bayesian literature to derive an approximate inference. These methods can be roughly classified into two broad categories: *stochastic* (sampling-based) and *deterministic* strategies. In the former the approximation is carried out simulating from the posterior distribution and performing inference on the resulting sample. This class includes the Monte Carlo Markov Chain (MCMC) methods (Chen et al. (2000) and Robert & Casella (2004)), such as the Gibbs sampling (GS) or the Metropolis-Hastings algorithm, and the more recent Approximate Bayesian Computation (see the seminal works by Rubin et al. (1984) and Tavaré et al. (1997)), that allows to bypass the analytic evaluation of the likelihood function for complex models. In the latter the basic idea is the replacement of the true posterior with a more tractable distribution. A classical example is given by the Laplace method, which considers the normal approximation located at the maximum *a posteriori* (MAP) estimate, and related developments in the context of latent Gaussian models, such as the Integrated Nested Laplace approximation described in Rue et al. (2009). The Variational Bayesian methods, instead, identify the analytic approximation as the member of a preliminarily chosen parametric family such that the Kullback-Leibler (KL) divergence of the original posterior from it is minimum (Jordan et al. (1999), MacKay (2003) and Beal (2003)). Reversing the role of the two distributions in the definition of the KL divergence, one has a different deterministic technique called Expectation-Propagation algorithm (Minka, 2001). The EP is, in turn, a special case of the more general Power EP introduced by Minka (2004), proved to be equivalent to the minimization of the α-divergence in place of the KL measure.

This chapter contributes with a brief account on the existing Bayesian ranking models and the proposal of a Bayesian mixture of PL to model partial rankings in the presence of a group structure. As revealed by the overview in Section 5.1, most of the above approximating techniques have been explored and implemented also for the Bayesian estimation of ranking models. In the subsequent sections we will focus on the recent Bayesian contribution by Caron & Doucet (2012) and, in line with this work, we will limit our attention to the inference achieved via the MAP estimation and the GS procedure.

5.1 Review of Bayesian modeling for ranked data

One of the first Bayesian inferential analysis appeared in the ranking literature concerns the ordered statistics models. Yao & Böckenholt (1999) discuss the difficulties related to the estimation of the TM parameters, due in particular to the evaluation of the likelihood, and show that the GS can provide an efficient answer to high-dimensional integration thanks to the completion with the latent scores in the model specification. Philip (2000) enlarges further such an approach accounting for the introduction of individual covariates in the TM through a linear regression framework. In this regard, Johnson & Kuhn (2013) provide JAGS code (Plummer et al., 2003) for the implementation of the Bayesian TM. The GS is applied also in the Bayesian setting for DB models described by Gupta & Damien (2002). The authors exploit an equivalence relation on $\mathcal{S}_K$ to facilitate the prior specification for the modal sequence. In particular, the suggested class of prior distributions makes use of the Hausdorff distance for subgroups to return constant probabilities over the equivalence classes. Stagewise ranking models, instead, are considered in Guiver & Snelson (2009), who derive a deterministic approach to perform Bayesian inference on the PL. Their method is based on the Power EP algorithm, which simplifies the treatment of the annoying denominator of the PL likelihood by taking its reciprocal. As shown by Caron & Doucet (2012), an effective Bayesian PL estimation can be achieved also by means of a suitable data augmentation approach. We review their work in great detail in the next section, since it represents the starting point of our proposal.

Several Bayesian contributions to the ranking literature aim also at addressing the issue of model-based clustering. Gormley & Murphy (2009) construct a Bayesian *grade of membership model* (GoM) for ranked data to derive a soft clustering of the sample units. Soft-clustering means that each subject belongs to model-fitted groups probabilistically, rather than in a mutually exclusive way, with individual membership degrees called *GoM scores* or *mixed-membership parameters*. The crucial assumption underlying their GoM model is that mixed-membership parameters operate at each rank assignment step t rather than at the overall ranking level, as typically postulated in a standard mixture model. In particular, the probability that a generic sample unit s selects a certain item at the t-th stage is taken as the average of the same probability under each subpopulation g weighted with the individual GoM scores $\underline{\omega}_s = (\omega_{s1}, \ldots, \omega_{sG})$:

$$P(\pi_s^{-1}|\underline{p}, \underline{\omega}_s) = \prod_{t=1}^{K} \sum_{g=1}^{G} \omega_{sg} \frac{p_{g\pi_s^{-1}(t)}}{\sum_{\nu=t}^{K} p_{g\pi_s^{-1}(\nu)}}.$$

The GoM is estimated by Gormley & Murphy (2009) through a Metropolis-within-Gibbs algorithm. As further attempt to account for sample heterogeneity, one can mention the recent work by Meilă & Chen (2010), where the mixture is essentially induced placing a Dirichlet process as prior specification for the GMM parameters. The resulting model

$$\begin{cases} F \sim \mathrm{DP}(\alpha, P^0(\sigma, \underline{\lambda}|\nu, r)), \\ (\sigma_s, \underline{\lambda}_s) \sim F, \\ \pi_s|\sigma_s, \underline{\lambda}_s \sim \mathrm{GMM}(\sigma_s, \underline{\lambda}_s) \end{cases} \tag{5.1.1}$$

is referred to as *Dirichlet process mixture model* (DPMM) and represents an adaptation of the general modeling framework introduced by Lo et al. (1984) to address ranking data analysis. The authors describe in great detail an efficient GS scheme to conduct inference on (5.1.1). Since distribution functions sampled from a Dirichlet process are almost surely discrete, the above DPMM can be interpreted as a countably infinite mixture model. See Neal (2000) for a more transparent equivalent formulation of (5.1.1) as the limit for $G \to \infty$ of a Bayesian finite mixture model. The adopted base distribution $P^0(\sigma, \underline{\lambda}|\nu, r)$ is proved by Meilă & Bao (2008) to be the conjugate prior for the GMM. The hyperparameter α, modulating the initial confidence in P^0, affects the granularity of the mixture. In the present setting, in fact, the number of clusters is a parameter to be inferred, making the DPMM a useful tool in those situations where the number of groups is not *a priori* known. See, for example, Ali et al. (2010) for an application of (5.1.1) to the CAO data set. Nonparametric Bayesian clustering via DPMM has been recently proposed also by Caron et al. (2012) and Caron et al. (2014). In this series of works the generative model for partial rankings is the extended version of the Plackett-Luce model for infinite rankings, originally presented in Caron & Teh (2012).

5.2 Bayesian inference for the Plackett-Luce model

In this section we give an outline of the Bayesian approach recently proposed by Caron & Doucet (2012) to make inference on the PL parameters in the case of homogeneous population. In their work the PL parametric approach is referred to as a model for multiple comparisons, to stress the contrast with the BT for pairwise comparisons.

Let $\underline{\pi}^{-1} = \{\pi_1^{-1}, \ldots, \pi_N^{-1}\}$ be a random sample drawn from a PL, consisting of N partial top-n_s orderings where n_s specifies the number of top positions expressed by unit s in $\pi_s^{-1} = (\pi_s^{-1}(1), \ldots \pi_s^{-1}(n_s))$. Hence, in what follows we will consider observed sequences with unit-dependent lengths, where a full ordering corresponds to both the special cases $n_s = K - 1$ and $n_s = K$. In fact, once $K - 1$ items have been ranked, the last position is automatically determined. Caron & Doucet (2012) adopt an extended PL definition as sampling distribution, that is

$$\mathbf{P}_{\mathrm{PL}}(\pi_s^{-1}|\underline{p}) = \prod_{t=1}^{n_s} \frac{p_{\pi_s^{-1}(t)}}{\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}}. \tag{5.2.1}$$

The PL formulation given in (5.2.1) is the one employed in Hunter (2004) to account for the possible presence of partial top rankings. It exploits the property of internal

consistency of the PL (Guiver & Snelson, 2009) and leads to a normalization of the p's at each stage which is s-dependent, in the sense that it is taken w.r.t. the items actually ranked in every specific ordinal sequence. This likelihood approach is different, for example, from the PL specification for top rankings described in the reference monograph Marden (1995) and in the series of works by Gormley & Murphy (2006, 2008a,b, 2009), where the support parameter rescaling is constant over observed rankings and contemplates the whole item set I. It is useful to know and keep this difference in mind, although in the applications limited to full rankings it becomes irrelevant. Moreover, in this specific inferential framework item parameters $\underline{p}$ are still understood as unknown positive quantities but with no explicit unit-sum constraint $\sum_{i=1}^{K} p_i = 1$. However, we continue to count a total of $K-1$ free parameters since the item supports are identified up to multiplication by a positive constant.

The crucial idea in Caron & Doucet (2012) which leads to an efficient Bayesian estimation consists in an augmentation step realized with the introduction of continuous latent variables in the model specification. More specifically, they complete the sampling space with unobservable variables $\underline{Y} = (Y_{st})$ for $s = 1, \dots, N$ and $t = 1, \dots, n_s$, associated to each entry of the observed matrix, such that

$$f(\underline{y}|\underline{\pi}^{-1}, \underline{p}) = \prod_{s=1}^{N} \prod_{t=1}^{n_s} f_{\text{Exp}}\left(y_{st} \middle| \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}\right), \tag{5.2.2}$$

where $f_{\text{Exp}}(\cdot;\lambda)$ denotes the Negative Exponential density function parameterized by the rate parameter λ. The parametric assumption (5.2.2) does not alter the marginal structure of the model but, as detailed shortly, entails decisive facilitations in deriving closed-forms for both the optimization and the GS algorithm. In this regard, notice that the rate parameters $\lambda_{st} = \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}$ of the Y's correspond to the sequential normalizations of the p's, that is the annoying terms of the PL likelihood. Once the model governing observed and latent variables is specified, a fully Bayesian approach requires the elicitation of the joint prior distribution for the unknown parameters $\underline{p}$. As prior specification Caron & Doucet (2012) assume that the support parameters are i.i.d. as a Gamma r.v. implying

$$f_0(\underline{p}|c,d) = \prod_{i=1}^{K} f_{\text{Ga}}(p_i|c,d),$$

that is the same prior employed in Guiver & Snelson (2009) indexed by the shape and the rate parameters. The choice of the Gamma is motivated by its conjugacy with the Gumbel distribution which, as recalled in Section 1.2.2, reduces the TM to the PL when employed as distribution of the scores. In this setting the complete

log-likelihood function is

$$\begin{aligned}
l_c(\underline{p}, \underline{y}) &= \log \mathbf{P}(\underline{\pi}^{-1}, \underline{y}|\underline{p}) = \log f(\underline{y}|\underline{\pi}^{-1}, \underline{p})\mathbf{P}(\underline{\pi}^{-1}|\underline{p}) \\
&= \log \left(\prod_{s=1}^{N} \prod_{t=1}^{n_s} \left(\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)} \right) e^{-y_{st} \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}} \times \prod_{s=1}^{N} \prod_{t=1}^{n_s} \frac{p_{\pi_s^{-1}(t)}}{\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}} \right) \\
&= \log \prod_{s=1}^{N} \prod_{t=1}^{n_s} \left(\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)} \right) e^{-y_{st} \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}} \frac{p_{\pi_s^{-1}(t)}}{\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}} \\
&= \log \prod_{s=1}^{N} \prod_{t=1}^{n_s} p_{\pi_s^{-1}(t)} e^{-y_{st} \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}} \\
&= \sum_{s=1}^{N} \sum_{t=1}^{n_s} \left(\log p_{\pi_s^{-1}(t)} - y_{st} \sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)} \right).
\end{aligned}$$

The crucial step in the previous computation is the elimination of the annoying denominator of the PL likelihood thanks to the assumption (5.2.2).

As first type of inference Caron & Doucet (2012) describe how to achieve the MAP estimate of the vector $\underline{p}$, i.e., the mode of the posterior distribution. In the presence of the above latent variable model, they construct an EM algorithm to optimize the posterior distribution where the objective function, differently from the canonical frequentist implementation, contemplates also the prior information $f_0(\underline{p}|c, d)$. Explicitly, the E-step of the EM algorithm is given by

$$\begin{aligned}
Q(\underline{p}, \underline{p}^*) &= \mathbb{E}[l_c(\underline{p}, \underline{y})|\underline{\pi}^{-1}, \underline{p}^*] + \log f_0(\underline{p}|c, d) \\
&\propto \sum_{s=1}^{N} \sum_{t=1}^{n_s} \left(\log p_{\pi_s^{-1}(t)} - \frac{\sum_{\nu=t}^{n_s} p_{\pi_s^{-1}(\nu)}}{\sum_{\nu=t}^{K} p^*_{\pi_s^{-1}(\nu)}} \right) + \sum_{i=1}^{K} ((c-1) \log p_i - d p_i),
\end{aligned} \tag{5.2.3}$$

where $\propto$ incorporates all additive/multiplicative terms not depending on $\underline{p}$. Setting noninformative hyperparameters values $c = 1$ and $d = 0$, expression (5.2.3) reduces to the minorizing auxiliary objective function of the MM algorithm described by Hunter (2004) and aimed at the MLE. Differentiating and equating to zero, at iteration $l + 1$ one obtains the following update of the PL parameters

$$p_i^{(l+1)} = \frac{c - 1 + \gamma_i}{d + \sum_{s=1}^{N} \sum_{t=1}^{n_s} \frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p^{(l)}_{\pi_s^{-1}(\nu)}}} \qquad i = 1, \ldots, K,$$

where

$$\delta_{sti} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(t), \ldots, \pi_s^{-1}(n_s)\}, \\ 0 & \text{otherwise} \end{cases}$$

and $\gamma_i = \sum_{s=1}^{N} u_{si}$ with

$$u_{si} = \begin{cases} 1 & \text{if } i \in \{\pi_s^{-1}(1), \ldots, \pi_s^{-1}(n_s)\}, \\ 0 & \text{otherwise.} \end{cases}$$

The quantity γ_i denotes the number of sample units who assigned a position to item i, whereas δ_{sti} is the indicator that item i appears in a position not better than t in the s-th partial ranking.

Caron & Doucet (2012) detail also the GS to draw a sample from the joint posterior distribution. The algorithm requires the identification of the *full-conditional distributions*, which are the distributions of a subset of the unobserved variables given all the remaining unknown quantities and the data. Notice that the augmented model representation (5.2.2) determines by construction the full-conditional of the Y's. Moreover, the full-conditionals for the support parameters are still members of the Gamma family thanks to the conjugate structure, specifically

$$\begin{aligned}\mathbf{P}(p_i|\underline{\pi}^{-1}, \underline{y}, p_{[-i]}) &\propto f_0(p_i|c,d)L_c(\underline{p},\underline{y}) \\ &\propto p_i^{c-1}e^{-dp_i}\prod_{s=1}^{N} p_i^{u_{si}}e^{-p_i\sum_{t=1}^{n_s}\delta_{sti}y_{st}} \\ &= p_i^{c+\gamma_i-1}e^{-p_i(d+\sum_{s=1}^{N}\sum_{t=1}^{n_s}\delta_{sti}y_{st})}.\end{aligned}$$

As usual $p_{[-i]}$ denotes the vector $\underline{p}$ of the support parameters without the i-th component. In conclusion, at iteration $l+1$ the GS proceeds iteratively as follows

- for $s = 1, \ldots, N$ and $t = 1, \ldots, n_s$ sample

$$Y_{st}^{(l+1)}|\underline{\pi}^{-1}, \underline{p}^{(l)} \sim \text{Exp}\left(\sum_{\nu=t}^{n_s} p^{(l)}_{\pi_s^{-1}(\nu)}\right),$$

- for $i = 1, \ldots, K$ sample

$$p_i^{(l+1)}|\underline{\pi}^{-1}, \underline{y}^{(l+1)} \sim \text{Ga}\left(c+\gamma_i, d+\sum_{s=1}^{N}\sum_{t=1}^{n_s}\delta_{sti}y_{st}^{(l+1)}\right).$$

The two estimation procedures can be conveniently combined employing the MAP solution as initialization of the chain in the MCMC simulation.

In the next sections we will propose the introduction of additional latent variables in order to suitably generalize the approach by Caron & Doucet (2012) to the finite mixture context.

5.3 Bayesian mixture of Plackett-Luce models

To our knowledge Bayesian inference of a finite PL mixture for partially ranked data has not been previously developed in the literature, although a wide variety of research contexts requires a model-based analysis that accounts for the presence of differential patterns in the observed sequences. Bayesian PL estimation appeared so far in the literature is either limited to the homogeneous case (see Guiver & Snelson (2009) and Caron & Doucet (2012)) or addressed in the mixture context for an infinite number of items by Caron et al. (2012) with a nonparametric method based on the DPMM.

5.3.1 Model and prior specification

Let $\underline{\pi}^{-1}$ be a random sample drawn from a PL mixture, in symbols

$$\pi_1^{-1}, \ldots, \pi_N^{-1}|\underline{p}, \underline{\omega} \overset{i.i.d.}{\sim} \sum_{g=1}^{G}\omega_g\mathbf{P}_{\text{PL}}(\pi_s^{-1}|\underline{p}_g). \tag{5.3.1}$$

Under the finite mixture setup, the contribution (5.3.1) of the generic observation to the observed-data likelihood can be obtained by marginalization w.r.t. a latent feature of the s-th sample unit, represented by the unknown group membership label

$$\underline{z}_s = (z_{s1}, \ldots, z_{sG}) | \underline{\omega} \overset{i.i.d.}{\sim} \text{Mult}(1, \underline{\omega} = (\omega_1, \ldots, \omega_G)).$$

In order to suitably generalize the data augmentation approach in Caron & Doucet (2012) within the finite mixture framework, we can account for the unobserved group structure acting on the underlying quantitative mechanism in this way

$$f(\underline{y}|\underline{\pi}^{-1}, \underline{z}, \underline{p}, \underline{\omega}) = \prod_{s=1}^{N} \prod_{t=1}^{n_s} f_{\text{Exp}} \left(y_{st} \Big| \prod_{g=1}^{G} \left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)} \right)^{z_{sg}} \right). \tag{5.3.2}$$

Additionally, we construct the joint prior distribution for the unknown parameters, in this case $\underline{p}$ and $\underline{\omega}$, postulating prior independence

$$f_0(\underline{p}, \underline{\omega}) = f_0(\underline{p}) f_0(\underline{\omega})$$

and, as in the homogeneous population context, we justify the choice of both priors with the conjugate structure. For the support parameters, in fact, we simply extend the initial distribution in Caron & Doucet (2012), defining for $g = 1, \ldots, G$ and $i = 1, \ldots, K$

$$p_{gi} \overset{i.}{\sim} \text{Ga}(c_{gi}, d_g),$$

where the assumption that within the same group the p's are equally distributed is relaxed. For the mixture weights, taking values in the $(G-1)$-dimensional simplex, we make the standard hypothesis

$$\underline{\omega} \sim \text{Dir}(\alpha_1, \ldots, \alpha_G).$$

5.3.2 MAP estimation

In the presence of the latent variables $\underline{y}$ and $\underline{z}$, the complete-data likelihood turns out to be

$$L_c(\underline{p}, \underline{\omega}, \underline{y}, \underline{z}) = \mathbf{P}(\underline{y}, \underline{\pi}^{-1}, \underline{z} | \underline{p}, \underline{\omega}) = f(\underline{y}|\underline{\pi}^{-1}, \underline{z}, \underline{p}, \underline{\omega}) \mathbf{P}(\underline{\pi}^{-1}, \underline{z} | \underline{p}, \underline{\omega}), \tag{5.3.3}$$

equal to the product of the full-conditional (5.3.2) by the standard complete-data likelihood of a mixture model specification without the vector $\underline{y}$. In order to simplify analytic steps, hereafter we will make frequent use of the following equivalence

$$\prod_{g=1}^{G} \left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)} \right)^{z_{sg}} = \sum_{g=1}^{G} z_{sg} \sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)},$$

which follows from the special binary characterization of the z's. With simple computations the factors in (5.3.3) can be rewritten so that a multinomial form in

the $\underline{z}$ appears as follows

$$\begin{aligned} f(\underline{y}|\underline{\pi}^{-1},\underline{z},\underline{p},\underline{\omega}) &= f(\underline{y}|\underline{\pi}^{-1},\underline{z},\underline{p}) \\ &= \prod_{s=1}^{N}\prod_{t=1}^{n_s}\prod_{g=1}^{G}\left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}\right)^{z_{sg}} e^{-\sum_{g=1}^{G} y_{st} z_{sg} \sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}} \\ &= \prod_{s=1}^{N}\prod_{t=1}^{n_s}\prod_{g=1}^{G}\left(\left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}\right) e^{-y_{st}\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}} \\ &= \prod_{s=1}^{N}\prod_{g=1}^{G}\left(\prod_{t=1}^{n_s}\left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}\right) e^{-\sum_{t=1}^{n_s} y_{st}\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}} \end{aligned}$$

and

$$\begin{aligned} \mathbf{P}(\underline{\pi}^{-1},\underline{z}|\underline{p},\underline{\omega}) &= \prod_{s=1}^{N}\mathbf{P}(\underline{\pi}_s^{-1},\underline{z}_s|\underline{p},\underline{\omega}) = \prod_{s=1}^{N}\mathbf{P}(\underline{\pi}_s^{-1}|\underline{z}_s,\underline{p})\mathbf{P}(\underline{z}_s|\underline{\omega}) \\ &= \prod_{s=1}^{N}\prod_{g=1}^{G}\left(\omega_g \prod_{t=1}^{n_s}\frac{p_{g\pi_s^{-1}(t)}}{\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}}. \end{aligned}$$

It follows that

$$\begin{aligned} L_c(\underline{p},\underline{\omega},\underline{y},\underline{z}) = &\prod_{s=1}^{N}\prod_{g=1}^{G}\left(\prod_{t=1}^{n_s}\left(\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}\right) e^{-\sum_{t=1}^{n_s} y_{st}\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}} \times \\ &\prod_{s=1}^{N}\prod_{g=1}^{G}\left(\omega_g \frac{\prod_{t=1}^{n_s} p_{g\pi_s^{-1}(t)}}{\prod_{t=1}^{n_s}\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}} \\ = &\prod_{s=1}^{N}\prod_{g=1}^{G}\left(\omega_g\left(\prod_{t=1}^{n_s} p_{g\pi_s^{-1}(t)}\right) e^{-\sum_{t=1}^{n_s} y_{st}\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}}\right)^{z_{sg}} \\ = &\prod_{s=1}^{N}\prod_{g=1}^{G}\left(\omega_g \prod_{i=1}^{K} p_{gi}^{u_{si}} e^{-p_{gi}\sum_{t=1}^{n_s}\delta_{sti}y_{st}}\right)^{z_{sg}}, \end{aligned}$$

where u_{si} and δ_{sti} are the binary indicators previously described. Indicating with $l_c(\underline{p},\underline{\omega},\underline{y},\underline{z})$ the complete-data log-likelihood, we have

$$l_c(\underline{p},\underline{\omega},\underline{y},\underline{z}) = \sum_{s=1}^{N}\sum_{g=1}^{G} z_{sg}\left(\log\omega_g + \sum_{i=1}^{K}\left(u_{si}\log p_{gi} - p_{gi}\sum_{t=1}^{n_s}\delta_{sti}y_{st}\right)\right).$$

The implementation of the EM algorithm to derive the MAP estimates requires the iterative maximization of the following objective function

$$Q((\underline{p},\underline{\omega}),(\underline{p}^*,\underline{\omega}^*)) = \mathbb{E}_{\underline{y},\underline{z}|\underline{\pi}^{-1},\underline{p}^*,\underline{\omega}^*}[l_c(\underline{p},\underline{\omega},\underline{y},\underline{z})] + \log f_0(\underline{p},\underline{\omega}),$$

where in this case the expectation is computed w.r.t. the joint distribution for the latent variables given by

$$\mathbf{P}(\underline{y},\underline{z}|\underline{\pi}^{-1},\underline{p},\underline{\omega}) = f(\underline{y}|\underline{\pi}^{-1},\underline{z},\underline{p},\underline{\omega})\mathbf{P}(\underline{z}|\underline{\pi}^{-1},\underline{p},\underline{\omega}). \tag{5.3.4}$$

Computing first the expectation w.r.t $\underline{y}$, one has

$$\sum_{s=1}^{N}\sum_{g=1}^{G} z_{sg}\left(\log\omega_g + \sum_{i=1}^{K}\left(u_{si}\log p_{gi} - p_{gi}\sum_{t=1}^{n_s}\frac{\delta_{sti}}{\prod_{g=1}^{G}\left(\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}\right)^{z_{sg}}}\right)\right).$$

For the expectation w.r.t $\underline{z}$ it is convenient to rewrite the fractional term as

$$\frac{\delta_{sti}}{\prod_{g=1}^{G}\left(\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}\right)^{z_{sg}}} = \frac{\delta_{sti}}{\sum_{g=1}^{G} z_{sg}\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}} = \sum_{g=1}^{G} z_{sg}\frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}}$$

and, noting that $\mathbf{P}(\underline{z}|\underline{\pi}^{-1}, \underline{p}, \underline{\omega})$ in the right-hand side of (5.3.4) is proportional to $\mathbf{P}(\underline{\pi}^{-1}, \underline{z}|\underline{p}, \underline{\omega})$, the E-step returns

$$\begin{aligned} Q((\underline{p},\underline{\omega}),(\underline{p}^*,\underline{\omega}^*)) = &\sum_{s=1}^{N}\sum_{g=1}^{G}\hat{z}_{sg}\left(\log\omega_g + \sum_{i=1}^{K}\left(u_{si}\log p_{gi} - p_{gi}\sum_{t=1}^{n_s}\frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}}\right)\right) \\ &+ \sum_{g=1}^{G}(\alpha_g - 1)\log\omega_g + \sum_{g=1}^{G}\sum_{i=1}^{K}((c_{gi}-1)\log p_{gi} - d_g p_{gi}), \end{aligned}$$

where

$$\hat{z}_{sg} = \frac{\omega^*_g \mathbf{P}_{\mathrm{PL}}(\pi_s^{-1}|\underline{p}^*_g)}{\sum_{g'=1}^{G}\omega^*_{g'}\mathbf{P}_{\mathrm{PL}}(\pi_s^{-1}|\underline{p}^*_{g'})}.$$

Differentiating w.r.t. each p_{gi} we obtain

$$\frac{\partial Q}{\partial p_{gi}} = \frac{\sum_{s=1}^{N}\hat{z}_{sg}u_{si}}{p_{gi}} - \sum_{s=1}^{N}\hat{z}_{sg}\sum_{t=1}^{n_s}\frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}} + \frac{c_{gi}-1}{p_{gi}} - d_g \tag{5.3.5}$$

and equating (5.3.5) to zero, the updating rule for the support parameters turns out to be

$$p_{gi} = \frac{c_{gi} - 1 + \hat{\gamma}_{gi}}{d_g + \sum_{s=1}^{N}\hat{z}_{sg}\sum_{t=1}^{n_s}\frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p^*_{g\pi_s^{-1}(\nu)}}}$$

for $g = 1, \ldots, G$ and $i = 1, \ldots, K$, where

$$\hat{\gamma}_{gi} = \sum_{s=1}^{N}\hat{z}_{sg}u_{si}.$$

The optimization of $Q((\underline{p},\underline{\omega}),(\underline{p}^*,\underline{\omega}^*))$ w.r.t. $\underline{\omega}$ is subject to the canonical constraint $\sum_{g=1}^{G}\omega_g = 1$ and requires the expression of the Lagrangian function $\mathcal{L}$. Discarding additive terms not depending on $\underline{\omega}$ we obtain

$$\mathcal{L} = \sum_{s=1}^{N}\sum_{g=1}^{G}\hat{z}_{sg}\log\omega_g + \sum_{g=1}^{G}(\alpha_g - 1)\log\omega_g - \lambda\left(\sum_{g=1}^{G}\omega_g - 1\right),$$

where λ denotes the Lagrange multiplier. Hence,

$$\frac{\partial \mathcal{L}}{\partial \omega_g} = 0 \quad \Rightarrow \quad \omega_g = \frac{\sum_{s=1}^{N} \hat{z}_{sg} + \alpha_g - 1}{\lambda}.$$

Exploiting the constraint, one has $\lambda = \sum_{g'=1}^{G} \alpha_{g'} - G + N$ and thus the estimation formula for the mixture weights is

$$\omega_g = \frac{\alpha_g - 1 + \sum_{s=1}^{N} \hat{z}_{sg}}{\sum_{g'=1}^{G} \alpha_{g'} - G + N} \qquad g = 1, \ldots, G.$$

In conclusion the EM reduces to the following iterative procedure:

Initialization: set starting values $\underline{p}^{(0)}, \underline{\omega}^{(0)}$ for the parameters to be estimated;

Computation: at iteration $l+1$, compute until convergence

- for $s = 1, \ldots, N$ and $g = 1, \ldots, G$

$$\hat{z}_{sg}^{(l+1)} = \frac{\omega_g^{(l)} \mathbf{P}_{\mathrm{PL}}(\pi_s^{-1} | \underline{p}_g^{(l)})}{\sum_{g'=1}^{G} \omega_{g'}^{(l)} \mathbf{P}_{\mathrm{PL}}(\pi_s^{-1} | \underline{p}_{g'}^{(l)})},$$

- for $g = 1, \ldots, G$

$$\omega_g^{(l+1)} = \frac{\alpha_g - 1 + \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)}}{\sum_{g'=1}^{G} \alpha_{g'} - G + N},$$

- for $g = 1, \ldots, G$ and $i = 1, \ldots, K$

$$p_{gi}^{(l+1)} = \frac{c_{gi} - 1 + \hat{\gamma}_{gi}^{(l+1)}}{d_g + \sum_{s=1}^{N} \hat{z}_{sg}^{(l+1)} \sum_{t=1}^{n_s} \frac{\delta_{sti}}{\sum_{\nu=t}^{n_s} p_{g\pi_s^{-1}(\nu)}^{(l)}}}.$$

When flat priors are employed in the analysis, corresponding to fixing $c_{gi} = 1$, $d_g = 0$ and $\alpha_g = 1$, such an estimation scheme coincides with the EMM algorithm in Gormley & Murphy (2006).

5.3.3 Gibbs sampling

The prior configuration described in Section 5.3.1, combined with the evidence provided by the data, leads to a direct posterior update of the hyperparameters and, hence, to a sampling scheme with parametric distributions which are simple to draw from. As in the homogeneous case, the derivation of the full-conditional distributions requires the complete-data likelihood $L_c(\underline{p}, \underline{\omega}, \underline{y}, \underline{z})$ detailed in the previous section. The full-conditionals of the latent component labels are easily derived noting that

$$\mathbf{P}(\underline{z} | \underline{\pi}^{-1}, \underline{y}, \underline{p}, \underline{\omega}) \propto L_c(\underline{p}, \underline{\omega}, \underline{y}, \underline{z}),$$

implying the following multinomial structure

$$\mathbf{P}(\underline{z}_s | \pi_s^{-1}, \underline{y}_s, \underline{p}, \underline{\omega}) \propto \prod_{g=1}^{G} \left(\omega_g \prod_{i=1}^{K} p_{gi}^{u_{si}} e^{-p_{gi} \sum_{t=1}^{n_s} \delta_{sti} y_{st}} \right)^{z_{sg}}.$$

The full-conditionals of the support parameters are Gamma with hyperparameters updated as follows

$$\begin{aligned}\mathbf{P}(p_{gi}|\underline{\pi}^{-1},\underline{y},\underline{z},p_{[-gi]},\underline{\omega}) &\propto f_0(p_{gi})L_c(\underline{p},\underline{\omega},\underline{y},\underline{z})\\ &\propto p_{gi}^{c_{gi}-1}e^{-d_g p_{gi}}\prod_{s=1}^{N}\left(p_{gi}^{u_{si}}e^{-p_{gi}\sum_{t=1}^{n_s}\delta_{sti}y_{st}}\right)^{z_{sg}}\\ &= p_{gi}^{c_{gi}+\gamma_{gi}-1}e^{-p_{gi}(d_g+\sum_{s=1}^{N}z_{sg}\sum_{t=1}^{n_s}\delta_{sti}y_{st})},\end{aligned}$$

where

$$\gamma_{gi}=\sum_{s=1}^{N}z_{sg}u_{si}$$

is the number of units assigned to cluster g who have ranked item i. Also the full-conditional of the mixture weights turns out to belong to a well-known parametric density, thanks to the conjugacy of the Dirichlet prior distribution with the multinomial model. In fact, one has

$$\begin{aligned}\mathbf{P}(\underline{\omega}|\underline{\pi}^{-1},\underline{y},\underline{z},\underline{p}) &\propto f_0(\underline{\omega})L_c(\underline{p},\underline{\omega},\underline{y},\underline{z}) \propto f_0(\underline{\omega})\mathbf{P}(\underline{z}|\underline{\omega})\\ &\propto \prod_{g=1}^{G}\omega_g^{\alpha_g-1}\times\prod_{s=1}^{N}\prod_{g=1}^{G}\omega_g^{z_{sg}}\\ &= \prod_{g=1}^{G}\omega_g^{\alpha_g+\sum_{s=1}^{N}z_{sg}-1}.\end{aligned}$$

Finally, the form of the full-conditionals for the y's is given by the assumption (5.3.2). In conclusion, the GS algorithm to approximate the joint posterior distribution $\mathbf{P}(\underline{z},\underline{y},\underline{p},\underline{\omega}|\underline{\pi}^{-1})$ consists in performing iteratively the following steps

Initialization: set the total number T of iterations and the starting values $\underline{y}^{(0)}$, $\underline{z}^{(0)}$, $\underline{p}^{(0)}$ and $\underline{\omega}^{(0)}$ (note that we need starting values only for $\underline{z}$ and $\underline{p}$);

Sampling: at iteration $(l+1)\leq T$, sample

- the mixture weights

$$\underline{\omega}^{(l+1)}|\underline{z}^{(l)}\sim\text{Dir}\left(\alpha_1+\sum_{s=1}^{N}z_{s1}^{(l)},\ldots,\alpha_G+\sum_{s=1}^{N}z_{sG}^{(l)}\right),$$

- for $s=1,\ldots,N$ and $t=1,\ldots,n_s$

$$Y_{st}^{(l+1)}|\pi_s^{-1},\underline{z}_s^{(l)},\underline{p}^{(l)}\sim\text{Exp}\left(\prod_{g=1}^{G}\left(\sum_{\nu=t}^{n_s}p_{g\pi_s^{-1}(\nu)}^{(l)}\right)^{z_{sg}^{(l)}}\right),$$

- for $g=1,\ldots,G$ and $i=1,\ldots,K$

$$p_{gi}^{(l+1)}|\underline{\pi}^{-1},\underline{y}^{(l+1)},\underline{z}^{(l)}\sim\text{Ga}\left(c_{gi}+\gamma_{gi}^{(l)},d_g+\sum_{s=1}^{N}z_{sg}^{(l)}\sum_{t=1}^{n_s}\delta_{sti}y_{st}^{(l+1)}\right),$$

- for $s = 1, \ldots, N$

$$\underline{z}_s^{(l+1)} | \pi_s^{-1}, \underline{y}_s^{(l+1)}, \underline{p}^{(l+1)}, \underline{\omega}^{(l+1)} \sim \text{Mult}\left(1, \left(m_{s1}^{(l+1)}, \ldots, m_{sG}^{(l+1)}\right)\right),$$

where

$$m_{sg}^{(l+1)} = \frac{\omega_g^{(l+1)} \prod_{i=1}^{K} (p_{gi}^{(l+1)})^{u_{si}} e^{-p_{gi}^{(l+1)} \sum_{t=1}^{n_s} \delta_{sti} y_{st}^{(l+1)}}}{\sum_{g'=1}^{G} \omega_{g'}^{(l+1)} \prod_{i=1}^{K} (p_{g'i}^{(l+1)})^{u_{si}} e^{-p_{g'i}^{(l+1)} \sum_{t=1}^{n_s} \delta_{sti} y_{st}^{(l+1)}}}.$$

5.4 Bayesian model comparison

In the estimation procedures previously described, the number G of groups is fixed *a priori*. When we perform inference on PL mixtures with alternative values of G, a method for discriminating among the competing models is needed. Due to the computational difficulties related to the marginal likelihood-based methods, such as the Bayes factor, for the selection of the number of groups we rely on two alternative Bayesian criteria, i.e., the Deviance Information Criterion (DIC) and the Bayesian Predictive Information Criterion (BPIC). The DIC was originally introduced in the fundamental work by Spiegelhalter et al. (2002) and is a very popular penalized measure of fitting for Bayesian model comparison, particularly useful when inference is carried out with an MCMC sampling scheme. Denoting with $D(\theta) = -2 \log L(\theta)$ the monotone transformation of likelihood known in the statistical theory as *deviance*, the original version of the DIC described by Spiegelhalter et al. (2002) has the following expression

$$DIC = \bar{D} + p_D.$$

The former term $\bar{D} = \mathbb{E}[D(\theta)|\underline{x}]$ is the posterior expected deviance, interpreted as a Bayesian overall measure of goodness-of-fit averaged w.r.t. the posterior distribution, whereas p_D is the *effective number of parameters* of the Bayesian model acting as penalty term. The penalty p_D is computed as the difference $\bar{D} - D(\tilde{\theta})$, where $D(\tilde{\theta})$ is the deviance evaluated at the single point estimate $\tilde{\theta}$. In the literature an alternative method to assess model complexity was proposed by Gelman et al. (2004), consisting in setting p_D equal to half the posterior variance of the deviance. Both computations of the effective number of parameters are justified by the asymptotic posterior distribution of the deviance, given by

$$D(\theta)|\underline{x} \sim D(\hat{\theta}_{ML}) + \chi^2_{p^*},$$

where $D(\hat{\theta}_{ML})$ is the ordinary frequentist deviance measure of the goodness-of-fit and p^* is the true number of parameters. The former version of the DIC estimates model complexity relying on the asymptotic posterior mean, whereas the latter on the asymptotic posterior variance. Once the MCMC sample is available, these quantities are easily approximated by the empirical mean and variance of the drawings from the posterior distribution of the deviance, that is the plugged-in values $D(\theta^{(l)})$. The large sample analyses presented in Section 5.7 and 5.8 motivate us to exploit the above asymptotic result and explore both DIC formulations, indicated respectively with DIC_1, based on $p_D = \bar{D} - D(\theta_{MAP})$, and DIC_2, based on $p_D = \mathbb{VAR}[D(\theta)|\underline{x}]/2$.

One aspect often debated on the DIC is its tendency to overfitting due to the double usage of the observed data. So, we additionally considered the BPIC suggested by Ando (2007), which penalizes the fitting measure $\bar{D}$ with $2p_D$.

The optimal model is identified with the one that minimizes the criterion.

5.5 Identifiability

When one adopts an MCMC simulation to derive approximate Bayesian inference of a mixture model, an annoying identifiability issue can affect the posterior sample. As emphasized by Koopmans & Reiersol (1950), identifiability is a very general and crucial issue related to any research context where a parametric modeling is assumed. The identifiability problem, in fact, concerns the mathematical structure postulated for the stochastic generating-data mechanism. Although the issue arises at theoretical level of the statistical analysis with the specification of a certain distributional law for the observations, the lack of identifiability prevents from conducting a conclusive inference on the parameters of interest, i.e., from deriving estimates in a univocal manner. Despite its theoretical and practical relevance, this subject is often neglected by practitioners. In the following two sections we limit ourselves to introduce some basic notions and the definition of a specific form of unidentifiability affecting mixture models, known as label-switching. Finally, in Section 5.7 and 5.8 we will discuss the implementation of our Bayesian PL mixture on both simulated and real data, combined with existing alternative methods to solve the label-switching issue.

Let $\mathcal{P} = \{P_\theta(\cdot) : \theta \in \Theta\}$ be a generic statistical model with probability distributions indexed by θ. Before giving the general definition of identifiability for a parametric model, it is useful to introduce the notion of observational equivalence for the elements of the parameter space Θ. Adopting the terminology in Rothenberg (1971) and Paulino & de Bragança Pereira (1994), two parameter configurations θ_1 and θ_2 are said to be *observational equivalent* (o.e.) if they determine the same distribution function for the data, that is

$$P_{\theta_1}(x) = P_{\theta_2}(x) \qquad x \in \mathcal{X}. \tag{5.5.1}$$

This implies that two o.e. parameter values are indistinguishable for any sample realization $\underline{x}$ because, by construction, the likelihood function turns out to be constant over these points, i.e., $L(\theta_1|\underline{x}) = L(\theta_2|\underline{x})$ for all $\underline{x}$. In this condition multiple points could maximize the likelihood and, basing only on the data, it would not be possible to identify a unique estimate solution. Definition (5.5.1) induces a equivalence relation on the parameter space Θ and, hence, a partition into equivalence classes composed of o.e. parameter points. A parametric model is *identifiable* if any pair of parameter values implies different sampling distributions or, in other words, if all the equivalence classes defined by (5.5.1) are made of singletons. Thus, lack of identifiability is an intrinsic feature of the model specification and must not be confused with a deficiency of the data which, as stressed by Paulino & de Bragança Pereira (1994), can discriminate only among equivalence classes but not within them.

Unidentifiability is often seen as a minor problem in the Bayesian approach where the introduction of an informative prior over Θ can further discriminate the param-

eter values and break down the symmetry of the likelihood over the equivalence classes. In this regard an important point should be stressed: in unidentifiability regime the precise measurement principle by Savage (1962) no longer applies, in favor of a more delicate and less clear updating process. This means that for unidentifiable models the impact of the prior on the inferential results does not vanish as the sample size increases. Hence, a careful assessment of the actual sample information in the Bayesian analysis of unidentifiable models is strongly recommended, for example by means of a sensitivity analysis (Gustafson, 2010).

Although Bayesian procedures described in Section 5.3.2 and 5.3.3 are effective tools to model prior knowledge on the parameters, in our analyses we do not contemplate informative priors. We are mainly interested in making a direct comparison with the MLE, taking advantage of assessing estimation uncertainty in a more natural and possibly less computational demanding way.

5.6 Label-switching

From the above definitions it follows that the presence of a single pair of o.e. parameter points is sufficient for a model to lack identifiability. However, as clarified by Paulino & de Bragança Pereira (1994), typical forms of unidentifibility are such that none of the equivalence classes is singular. Mixture modeling exemplifies very well this aspect.

Unidentifiability of a mixture model is due to the so-called *label-switching* phenomenon (LS), that reflects the arbitrary attribution of the indices $\{1, \dots, G\}$ to denote the mixture components. The application of a permutation $\tau \in \mathcal{S}_G$ of the G indices to a given parameter point, which corresponds to a relabeling of the latent classes, does not modify the resulting sampling distribution. Thus, in a mixture setting any parameter θ has $G! - 1$ o.e. configurations. Formally, let $\theta = ((\omega_1, \eta_1), \dots, (\omega_G, \eta_G))$ be the generic parameter vector indexing the mixture family, where η_g collects the parameters of the g-th group and ω_g is the corresponding weight. Using the notation by Jasra et al. (2005), we denote with τ a permutation of the group labels $\{1, \dots, G\}$ and with $\tau(\theta)$ the rearrangement of the parameter components such that $\tau(\theta) = ((\omega_{\tau(1)}, \eta_{\tau(1)}), \dots, (\omega_{\tau(G)}, \eta_{\tau(G)}))$. This formalism allows us to translate the LS in the following invariance condition

$$f(x|\theta) = \sum_{g=1}^{G} \omega_g f(x|\eta_g) = \sum_{g=1}^{G} \omega_{\tau(g)} f(x|\eta_{\tau(g)}) = f(x|\tau(\theta))$$

for all $x \in \mathcal{X}$ and $\tau \in \mathcal{S}_G$, implying

$$L(\theta|\underline{x}) = L(\tau(\theta)|\underline{x}) \tag{5.6.1}$$

for all $\underline{x}$ and $\tau \in \mathcal{S}_G$. Equality (5.6.1) explains the presence of $G!$ symmetric modes in the likelihood function independently on the realization $\underline{x}$ of the experiment. If the same permutation invariance is fulfilled by the prior distribution, no discrimination is produced within the equivalence classes and then the symmetry transfers also in the posterior distribution. In this situation, the marginal posterior distribution of each parameter is the same for all the mixture components and, hence, also the

posterior summaries coincide, nullifying the standard practice to use the posterior summaries as point estimates. As described in more detail shortly with an example on synthetic data, these aspects induce important difficulties in the application of sampling-based methods, such as the MCMC algorithms, for the Bayesian analysis of mixture models.

Different strategies have been proposed in the statistical literature to solve the LS issue in the MCMC analysis. Following the review by Jasra et al. (2005), they can be summarized in three classes: (i) introduction of artificial identifiability constraints, (ii) relabeling algorithms (RA) and (iii) employment of label-invariant loss functions.

The first approach consists in the elicitation of restrictions over the parameter space, typically order relations, which are satisfied by only one labeling τ of the mixture components. This action forces the equivalence classes to be singular so that LS ambiguity no longer persists. The pioneering work in this direction was Diebolt & Robert (1994), followed by the application of identifiability constraints in Richardson & Green (1997). Practical implications and criticisms related to this method are reviewed and discussed in Marin & Robert (2007) and Jasra et al. (2005). We simply note that in our multivariate setting the specification of artificial constraints can be very arduous.

The basic idea of the RA is the post-processing, that is the *ex-post* relabeling, of the raw MCMC simulations in order to make them lie in a unique posterior mode among the $G!$ possible modal regions. If the MCMC sample after the burn-in period has size T, the RA determines a total of T permutations for the rearrangement of each single drawing of the raw MCMC output. The Pivotal Reordering algorithm proposed by Marin et al. (2005), for example, switches the elements of each simulated value $\theta^{(l)}$ from the joint posterior distribution according to a permutation τ_l so that a certain distance from a target mode is minimized. The target mode plays the role of *pivot* and can be easily identified, for example, with the MAP solution. Note, in fact, that the MAP procedure is not affected by the LS issue because a single point estimate is returned by the optimization algorithm, without keeping track of the previous explorations of the parameter space. In this sense the posterior mode represents a naive answer to the LS, able to provide an unambiguously inference on the G subpopulations. On the other hand, it does not allow us to directly have any information about estimation uncertainty, which instead can be addressed by a sampling-based method. Moreover, MAP estimation fails in capturing "genuine" multimodality in the posterior distribution, that is the multimodality not induced by the LS, since its output is limited to the single best solution. The class of RA includes also the popular clustering-oriented method proposed by Stephens (2000), the Probabilistic RA by Sperrin et al. (2010), the Equivalence Classes Representatives technique suggested by Papastamoulis & Iliopoulos (2010) and the Data RA recently introduced by Rodríguez & Walker (2014). With the only exception of the Data RA, they are all implemented in a recently released R package, called `label.switching`, that we used in our Bayesian PL mixture applications.

We conclude this brief review of the solutions for the LS mentioning the strategy based on the decision theoretic approach, where meaningful Bayesian estimates in presence of LS are obtained with the minimization of the posterior expectation of label-invariant loss functions. We do not go into further details of such an approach

Table 5.1. Population scenarios considered in the simulation study and corresponding inferential results from the MAP estimate and the GS procedure initialized with random starting values. Posterior means were computed on both the raw and the relabeled MCMC samples. For Scenario 3 the GS was initialized also with the MAP solution, whose results are highlighted in grey.

	Scenario 1							Scenario 2							Scenario 3						
	g	$\hat{\omega}_g$	$\hat{\sigma}_g^{-1}$	$\hat{p}_{g1}$	$\hat{p}_{g2}$	$\hat{p}_{g3}$	$\hat{p}_{g4}$	g	$\hat{\omega}_g$	$\hat{\sigma}_g^{-1}$	$\hat{p}_{g1}$	$\hat{p}_{g2}$	$\hat{p}_{g3}$	$\hat{p}_{g4}$	g	$\hat{\omega}_g$	$\hat{\sigma}_g^{-1}$	$\hat{p}_{g1}$	$\hat{p}_{g2}$	$\hat{p}_{g3}$	$\hat{p}_{g4}$
True value	1	.700	(1, 2, 3, 4)	.700	.200	.080	.020	1	.700	(1, 2, 3, 4)	.700	.200	.080	.020	1	.700	(1, 2, 3, 4)	.700	.200	.080	.020
	2	.300	(4, 3, 2, 1)	.040	.120	.240	.600	2	.300	(1, 4, 3, 2)	.600	.040	.120	.240	2	.300	(1, 2, 4, 3)	.550	.300	.030	.120
MAP	1	.702	(1, 2, 3, 4)	.706	.205	.071	.018	1	.731	(1, 2, 3, 4)	.702	.209	.069	.020	1	.961	(1, 2, 3, 4)	.654	.252	.061	.033
	2	.298	(4, 3, 2, 1)	.049	.123	.261	.567	2	.269	(1, 4, 3, 2)	.554	.028	.101	.317	2	.039	(1, 4, 3, 2)	.499	.002	.003	.496
Raw GS	1	.699	(1, 2, 3, 4)	.699	.209	.073	.019	1	.732	(1, 2, 3, 4)	.695	.211	.072	.022	1	.515	(1, 2, 4, 3)	.569	.251	.079	.101
	2	.301	(4, 3, 2, 1)	.053	.129	.268	.551	2	.268	(1, 4, 3, 2)	.549	.031	.106	.315	2	.485	(1, 2, 4, 3)	.534	.274	.091	.102
															1	.731	(1, 2, 3, 4)	.623	.240	.071	.066
															2	.269	(1, 2, 4, 3)	.465	.283	.100	.152
Relabeled GS	1	.699	(1, 2, 3, 4)	.699	.209	.073	.019	1	.732	(1, 2, 3, 4)	.695	.211	.072	.022	1	.868	(1, 2, 3, 4)	.670	.231	.064	.035
	2	.301	(4, 3, 2, 1)	.053	.129	.268	.551	2	.268	(1, 4, 3, 2)	.549	.031	.106	.315	2	.132	(1, 2, 4, 3)	.433	.293	.106	.168
															1	.887	(1, 2, 3, 4)	.670	.231	.064	.035
															2	.113	(1, 2, 4, 3)	.418	.293	.106	.183

but the interested reader can refer to the fundamental works by Celeux et al. (2000) and Hurn et al. (2003) for both the theoretical and practical aspects related to the application of the label-invariant loss functions.

5.7 Simulation study

As first evaluation of the estimation procedures proposed in Section 5.3.2 and 5.3.3 we applied them in a simulation study, paying special attention to the behavior of the GS regarding the exploration of the parameter space and the occurrence of LS. For this purpose we simulated a sample of $N = 300$ complete orderings of length $K = 4$ from a 2-component PL mixture under three different population scenarios. True parameter values are reported in the upper-panel of Table 5.1 and describe bimodal populations with varying Kendall distances between the two modes, specifically equal to $d_K = 6$ in Scenario 1, $d_K = 3$ in Scenario 2 and $d_K = 1$ in Scenario 3. Adopting uninformative prior densities with hyperparameters equal to $c_{gi} = 1$, $d_g = .001$ and $\alpha_g = 1$, we performed for each simulated data set the MAP estimation through the EM algorithm and the GS initialized with random starting values. For the third scenario we considered also the initialization of the GS with the MAP estimate. We run the sampling algorithm for a total of 22000 iterations but, as usual practice for this approach, we discarded the first 2000 drawings (burn-in period) to avoid undesirable bias of the estimation results depending on the starting points. The LS problem did not take place at all in the GS application to the first two scenarios because the sampler explored only one of the two symmetric modes. In general, in fact, it is more likely that the LS occurs when the mixture components are not well-separated. This aspect confirms a well-known feature of the GS, unable to efficiently move over the parameter space and capture in full the multimodal profile of the posterior distribution. For the cases of artificial multimodality induced by the LS this behavior is paradoxically convenient because it allows straightforward inference. As remarked by Celeux

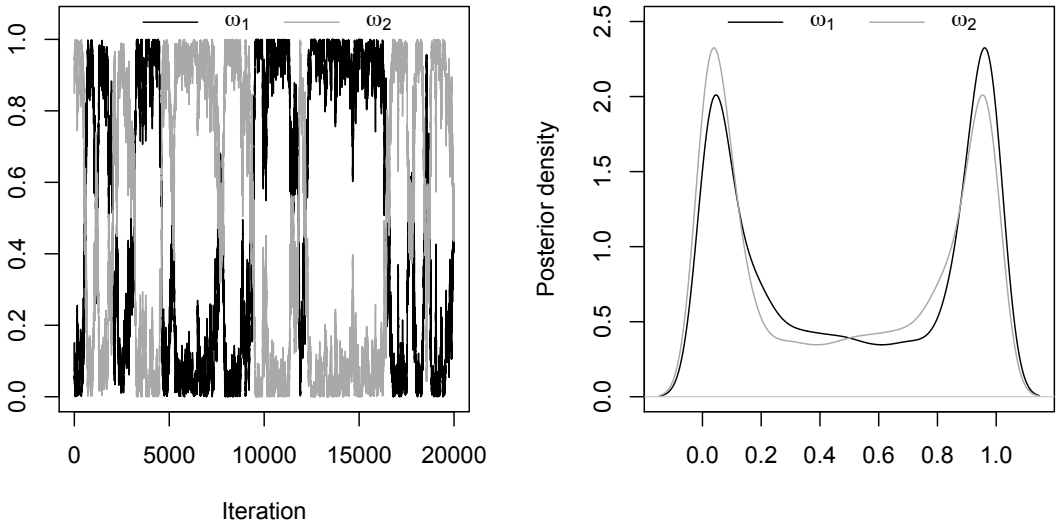

Figure 5.1. Traceplots and posterior marginal densities for the weights of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario.

et al. (2000), in fact, the visit of different symmetric modal regions is unnecessary from an inferential point of view, since the artificial peaks equivalently characterize the sampling distribution. On the other hand genuine multimodality, resulting for example from alternative representations suggested by the data, would not benefit from the inertia of the sampler in leaving a local mode. In this situation the MCMC sample could provide a partial picture of the posterior distribution and a consequent underestimation of the inferential uncertainty. Obviously, an inferential device able to distinguish between redundant and relevant information is desirable (see Grün & Leisch (2009) for a proposal in this context). However, the absence of LS for Scenario 1 and 2 allows us to directly derive meaningful GS estimates, which are indeed very close to the MAP results and in turn to the true parameter values, especially for the case of maximum distance between the two mixture components (Table 5.1).

In the GS analysis of the third scenario the effects of the LS are evident. They manifest with step-like configurations of the traceplots of the parameters indicating transitions of the sampler from one mode to another, see Figures 5.1 and 5.2. In particular, the traceplots point out that almost half of the chain is affected by LS, leading to marginal posterior densities which pretty much overlap, as shown in Figure 5.1 and 5.3, and yielding very similar empirical marginal means (Table 5.1). The LS still occurs when the MCMC chain is initialized with the MAP estimation but at a reduced extent. The dependence on the initialization reveals the instability of the LS occurrence for different runs of the MCMC algorithm.

The LS phenomenon clearly invalidates the use of posterior summaries based on the crude MCMC samples as meaningful estimates. For this reason, we are interested in assessing how alternative relabeling methods act in addressing the LS issue in the third scenario. The raw MCMC output has been post-processed using the functions implemented in the R package `label.switching`. The check of the traceplots reveals a good performance of all methods in removing the artificial

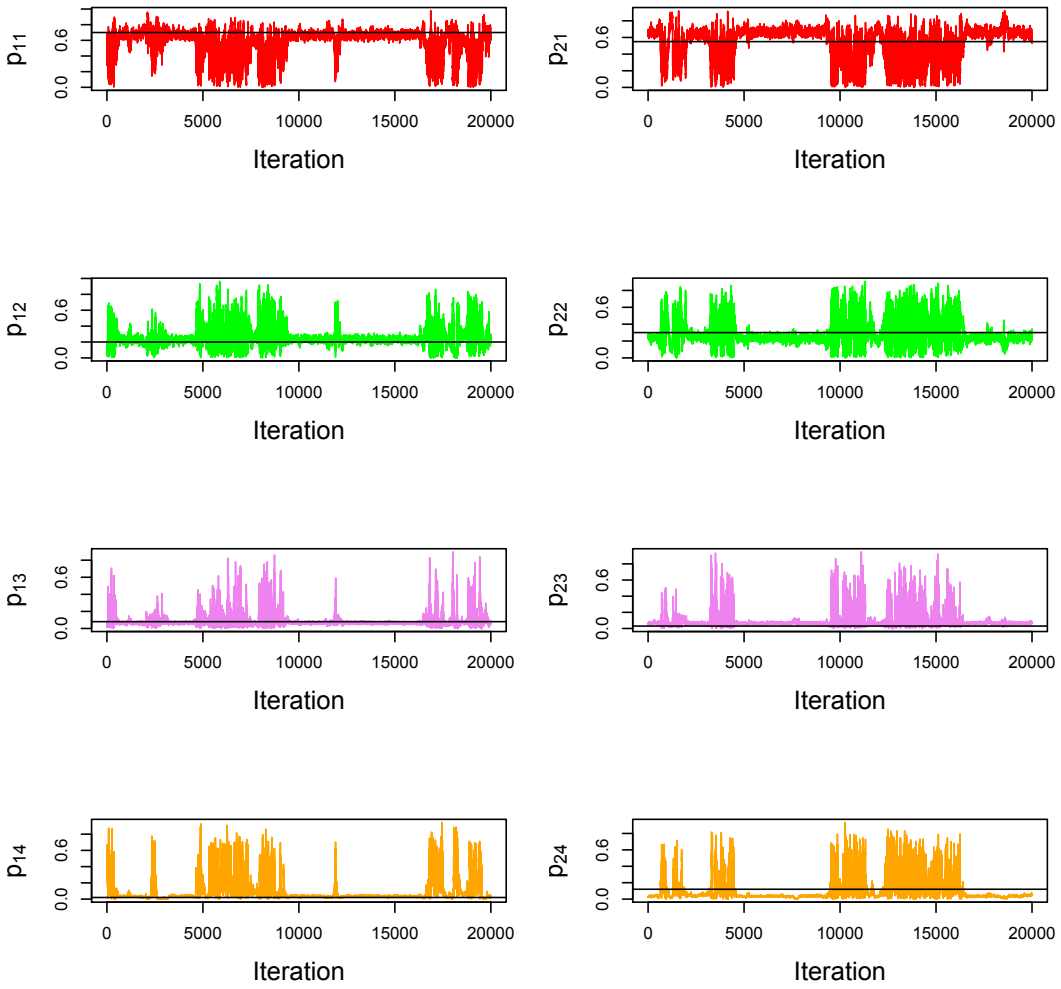

Figure 5.2. Traceplots for the support parameters of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario. Solid lines indicate the true parameter values.

multimodality. All strategies consistently rearranged 49% of the drawings and returned very similar results in terms of adjusted estimates. Posterior means derived specifically from the application of the algorithm proposed by Stephens (2000) are shown on the bottom-right of Table 5.1. Analogous considerations are valid for the results relative to the MCMC chain initialized with the MAP but, in this case, only 21% of the simulations has been permuted (see results in Table 5.1 highlighted in grey). As consequence of the adjacency of the two mixture components in the third scenario, we note that the closeness of the adjusted GS estimates to the true values is slightly reduced than that observed in both Scenario 1 and 2, although the actual order of the support parameters within each group is fully recovered (Table 5.1). When the two mixture components considerably overlap, in fact, it is more difficult to reconstruct the actual group memberships of the sample units and this can have negative effects on the final estimates. In this regard, the performance of the GS turns out to be better than the MAP procedure, which indeed corresponds to the MLE, since the latter completely fails in inferring the minor mixture component.

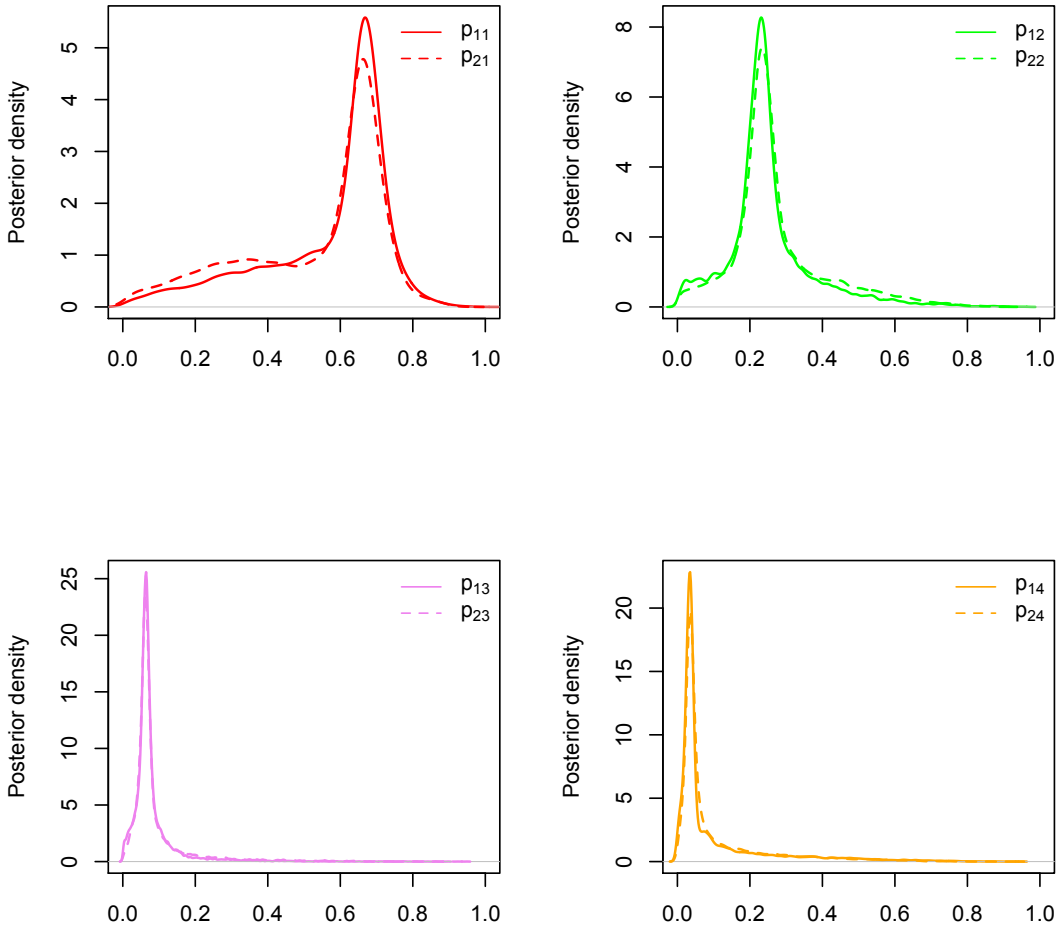

Figure 5.3. Posterior marginal densities for the support parameters of the 2-component PL mixture resulting from the raw MCMC sample for the third population scenario.

The posterior mode, in fact, exhibits the tendency of the optimization algorithm to privilege the homogenous model, i.e., to allocate the sample units into the major mode, returning a very poor estimates for the minor subpopulation. This aspect strongly affects also the criteria for Bayesian model comparison. Table 5.2 shows that in the third scenario both the DIC and BPIC are minimized by the homogeneous model, suggesting a mild evidence of heterogeneity induced by the strong similarity of the two generating mixture components. For Scenario 1 and 2, instead, all the criteria recognize the true number of groups.

5.8 Bayesian PL mixture for the HPQ data

In this section we illustrate the Bayesian PL mixture model with an application to the real data from the Hamburger Preparation Quiz (HPQ), carried out as part of the Menu Census Survey organized by the Market Research Corporation of America during the period March 1996−February 1997. The HPQ data set collects $N = 594$ complete rankings of $K = 5$ hamburger cooking methods ordered by the respondents according to their taste preferences. Hamburger preparation types, labeled from 1 to

Table 5.2. DIC and BPIC values for the Bayesian PL mixtures with varying number of components fitted to the simulated data from three population scenarios. Optimal values of the criteria are indicated in bold.

	Scenario 1				Scenario 2				Scenario 3			
G	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$
1	1769.4	1769.4	1772.4	1772.4	1507.7	1507.7	1510.7	1510.7	**1343.1**	**1343.1**	**1346.1**	**1346.1**
2	**1574.0**	**1573.9**	**1581.0**	**1580.9**	**1460.1**	**1459.9**	**1467.2**	**1466.9**	1348.5	1344.2	1357.1	1348.5
3	1576.8	1575.7	1589.6	1587.4	1462.9	1460.6	1472.9	1468.1	1356.0	1345.3	1372.6	1351.1

Table 5.3. DIC and BPIC values for the Bayesian PL mixtures with varying number of components fitted to the HPQ data. Optimal values of the criteria are indicated in bold.

G	DIC_1	DIC_2	$BPIC_1$	$BPIC_2$
1	4764.31	4764.21	4768.36	4768.17
2	3905.99	3907.50	3913.88	3916.89
3	3725.97	3729.36	3741.46	3748.24
4	3644.44	3651.28	3667.13	3680.82
5	**3590.88**	**3587.77**	**3624.28**	**3618.06**
6	3593.73	3602.82	3645.57	3663.76
7	3598.60	3648.70	3668.07	3768.27

5, are respectively: rare, medium-rare, medium, medium-well and well-done. The same data set has been previously analyzed by Gormley & Murphy (2010), who evaluated the possible information contribution provided by the inclusion of socio-demographic covariates in the mixture setting, and by Bao & Meilă (2008), where sample heterogeneity was investigated by means of a nonparametric approach. In our Bayesian analysis we first set uninformative hyperparameters values for the prior specification ($c_{gi} = 1$, $d_g = .001$ and $\alpha_g = 1$) and implemented the EM algorithm to obtain the MAP estimates for the PL mixtures, with a number of components varying from $G = 1$ up to $G = 7$. Obviously, flat priors reduces this approach to MLE but, in order to gain information on the uncertainty associated to the Bayesian parameter estimates, we subsequently employed the MAP solutions to initialize the GS procedure. The marginal traceplots revealed the presence of LS in the posterior MCMC samples (results not shown). So, as in the simulation study, alternative relabeling strategies have been applied to the GS outputs. A total of 15000 iterations, after a 5000 burn-in period, has been considered adequate for meaningful posterior inference subsequently to a positive graphical inspection of the algorithm convergence, supported by the good mixing of the traceplots and the weak sample autocorrelation.

DIC and PBIC values for the models fitted to the HPQ data are shown in Table 5.3 and consistently indicate the Bayesian PL mixture with $G = 5$ components as the best fitting model. Corresponding parameter estimates obtained from the MAP procedure and from the posterior means of the relabeled MCMC samples are shown in Table 5.4. Support parameter estimates are also represented via mosaic plots

Table 5.4. MAP estimates and posterior means of the relabeled MCMC samples for the best Bayesian PL mixture fitted to the HPQ data. Ordered sequences indicate the component-specific modal profiles, whereas posterior standard deviations are shown in parentheses.

Estimation	g	$\hat{\omega}_g$	$\hat{\sigma}_g^{-1}$	$\hat{p}_{g1}$	$\hat{p}_{g2}$	$\hat{p}_{g3}$	$\hat{p}_{g4}$	$\hat{p}_{g5}$
MAP	1	.412	(5,4,3,2,1)	.000	.000	.008	.255	.736
	2	.238	(3,2,4,5,1)	.006	.262	.594	.132	.006
	3	.013	(2,4,1,5,3)	.000	.988	.000	.012	.000
	4	.177	(4,3,5,2,1)	.000	.002	.378	.613	.007
	5	.160	(2,1,3,4,5)	.316	.454	.220	.009	.001
Relabeled GS	1	.389 (.02)	(5,4,3,2,1)	.000 (<.01)	.000 (<.01)	.007 (<.01)	.245 (.03)	.747 (.03)
	2	.234 (.03)	(3,2,4,1,5)	.008 (<.01)	.262 (.04)	.594 (.04)	.129 (.03)	.007 (<.01)
	3	.025 (<.01)	(2,4,1,5,3)	.136 (.06)	.384 (.13)	.028 (.03)	.351 (.10)	.102 (.06)
	4	.197 (.03)	(4,3,5,2,1)	.001 (<.01)	.010 (<.01)	.356 (.05)	.596 (.05)	.037 (.02)
	5	.155 (.02)	(2,1,3,4,5)	.319 (.05)	.450 (.04)	.219 (.04)	.012 (<.01)	.001 (<.01)

in Figure 5.4. It is interesting to compare the optimal Bayesian PL mixture with the maximum likelihood inference performed by Gormley & Murphy (2010) which, in the present uninformative prior setting, is very similar to our MAP estimates. The four main clusters recognized by our Bayesian model essentially agree with those pointed out by Gormley & Murphy (2010), both for the estimated size and for the group-specific preference patterns. Nevertheless, some differences can be highlighted. The final Bayesian model turns out to be more parsimonious because it identifies a single relevant component (the first one in Table 5.4) to represent those sample units who strongly prefer the well-done cooking type (item 5) and, as second best choice, the medium-done hamburger (item 4). Another difference concerns the group with smaller size, which is labeled as the third cluster in Table 5.4. In Gormley & Murphy (2010) this component is characterized by the exclusive preference for the medium-rare hamburger (item 2) and a substantial indifference towards the remaining alternatives. The GS estimates, instead, describe a more assorted liking profile with larger support to both item 2 and 4. The last choice, with very low estimated support equal to .028, is the medium cooking (item 3). Compared to the other clusters, this one exhibits a peculiar pattern since most of the support is not placed on contiguous/similar levels of hamburger doneness.

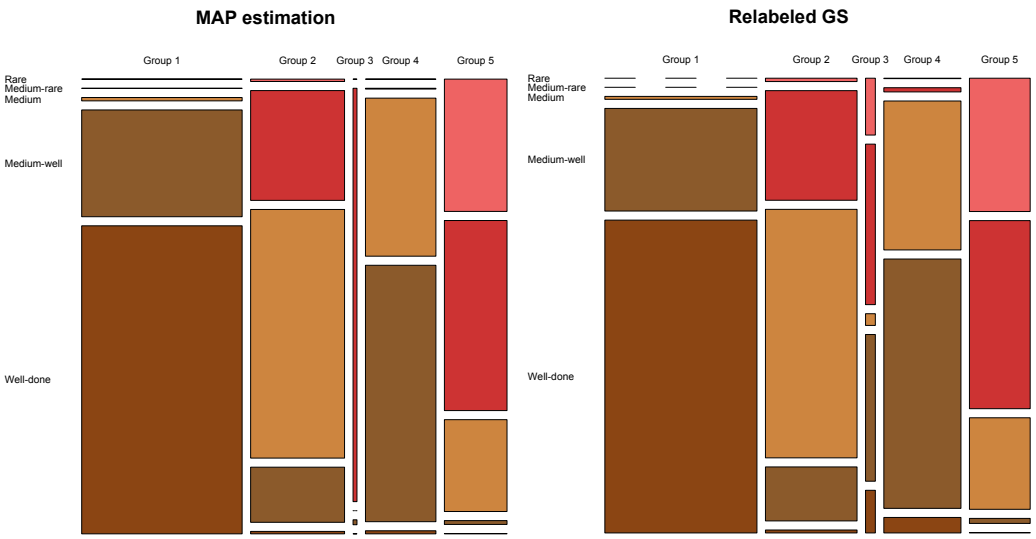

Figure 5.4. Mosaic plots for the support parameter estimates of the best Bayesian PL mixture fitted to the HPQ data. Bar widths are proportional to the estimated group weights.

Concluding remarks and future developments

This thesis has addressed the problem of parametric modeling for ranking data analysis. We have contributed some original extensions of the popular and widely-used Plackett-Luce model, considering different directions and inferential frameworks. As a first proposal we constructed the novel Extended Plackett-Luce model relaxing the standard assumption of forward ranking elicitation and detailed its estimation in the MLE approach. We verified the usefulness of the EPL with a successful application to the real LFPD data set from a bioassay experiment, comparing its performance to alternative and more standard probability distributions for rankings. Specifically, taking into account the heterogeneous origin of the sample units, we considered several parametric models in a mixture model setting. Inferential results of our mixture modeling approach revealed a good capability of the absorbance rankings to fit heterogeneous and wildly fluctuating binding data as well as a good accuracy in discriminating the actual disease state. Interestingly, an almost uniform component has been estimated from the data. Unlike previous applications in the literature, where the uniform component was introduced to fit outliers/untypical observations, for the LFPD data such a component has a precise interpretation in characterizing a subgroup of healthy patients. The utility of ranking-based analysis for epitope mapping experiments is reinforced by the possibility to partially overcome difficulties related to the choice of the preliminary normalization needed for the raw quantitative absorbance profiles. Additionally, the fitted model turned out to be more parsimonious than alternative quantitative analyses for the present multivariate setting and exhibited an interesting interpretation, unaffected by *ad-hoc* monotone pre-processing transformations of the original raw data. Hence, even when quantitative data are available in a bioassay experiment, statistical analysis of underlying ordinal information may provide a useful and more robust tool for the description of outcomes. Cluster-specific parameter estimates, characterizing groups of patients, are very useful to construct epitope mapping profiles. These can identify protein fragments whose binding can be related to the disease development and help detect spots relevant for possible classification/prediction purposes. The significantly improved fit obtained with the more general EPL class could suggest the absence of a natural and *a priori* known reference order of the binding mechanism and, in any case, the estimated reference order allowed to better capture the discriminant information of all positions.

The second proposal to ranking modeling is based on a further extension of the

EPL combining it with a well-established PL extension called Benter model. We integrated the two different PL generalizations in a wider parametric family named Benterized Extended Plackett-Luce model and solved the related MLE issues in both the homogeneous population case and the finite mixture context.

As a last contribution we have introduced the Bayesian mixture of Plackett-Luce models and described efficient algorithms to conduct approximate inference. Our Bayesian approach can be seen as a direct and natural extension of the Bayesian inference on multiple comparisons recently proposed in the literature, aimed at accounting for unobserved sample heterogeneity. At the same time, this contribution can be interpreted as the generalization within the Bayesian paradigm of the PL mixture, whose MLE has been achieved with a hybrid Expectation-Maximization algorithm and implemented for the analysis of ranking data from several preference and political studies. In the Bayesian analysis we additionally faced with the problem of label switching, that can affect the MCMC procedure and complicate the interpretation of inferential results, with the application of several relabeling strategies. The practical relevance of the proposed Bayesian PL mixture has been assessed with both a simulation study and the identification of group patterns in the real HPQ data set, concerning taste preferences of respondents towards different hamburger cooking styles.

A natural direction of development for further work could be the implementation of the EPL mixture model in the Bayesian framework, in order to allow the incorporation of pre-experimental information in the ranking analysis. This extension could benefit from the conjugacy of the PL with the Gamma prior distribution, already exploited in Guiver & Snelson (2009), Caron & Doucet (2012) as well as in our Bayesian proposal. A possible device to achieve this goal could be the addition in the Gibbs sampling scheme described for the PL mixture of a Metropolis-Hastings step to simulate the discrete reference order based, for example, on a random walk as proposal distribution.

Bibliography

Ali, A. & Meilă, M. (2012). Experiments with Kemeny ranking: what works when? *Math. Social Sci.* 64 28–40.

Ali, A., Murphy, T. B., Meilă, M. & Chen, H. (2010). Preferences in college applications–a nonparametric Bayesian analysis of top-10 rankings. NIPS Workshop on Computational Social Science and the Wisdom of Crowds, 2010 (slides), URL `http://www.cs.umass.edu/~wallach/workshops/nips2010css/papers/ali.pdf`.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* 94 443–458.

Babington Smith, B. (1950). Discussion of Professor Ross's paper. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 12 53–56.

Bao, L. & Meilă, M. (2008). Clustering permutations by Exponential Blurring Mean-Shift algorithm. Tech. Rep. 524, Dept of Statistics, University of Washington.

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference*. Ph.D. thesis, University of London.

Benter, W. (1994). Computer based horse race handicapping and wagering systems: A report. In D. B. Hausch, V. S. Lo & W. T. Ziemba, eds., *Efficiency of Racetrack Betting markets*. Academic Press, 183–198.

Biernacki, C. & Jacques, J. (2013). A generative model for rank data based on insertion sort algorithm. *Computational Statistics & Data Analysis* 58 162–176.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B. G. (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46 373–388.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics* 32 213–232.

Bradley, R. A. (1984). Paired comparisons: some basic procedures and examples. In *Nonparametric methods*, vol. 4 of *Handbook of Statist.* North-Holland, 299–326.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs. *Biometrika* 39 324–345.

Busse, L. M., Orbanz, P. & Buhmann, J. M. (2007). Cluster analysis of heterogeneous rank data. In Z. Ghahramani, ed., *Proceedings of the 24th International Conference on Machine Learning – ICML 2007*. Omnipress, 113–120.

Caron, F. & Doucet, A. (2012). Efficient Bayesian inference for generalized Bradley-Terry models. *J. Comput. Graph. Statist.* 21 174–196.

Caron, F. & Teh, Y. W. (2012). Bayesian nonparametric models for ranked data. In *Neural Information Processing Systems – NIPS 2012*. 1529–1537.

Caron, F., Teh, Y. W. & Murphy, T. B. (2012). Bayesian nonparametric Plackett-Luce models for the analysis of clustered ranked data. ArXiv:1211.5037.

Caron, F., Teh, Y. W. & Murphy, T. B. (2014). Bayesian nonparametric Plackett-Luce models for the analysis of preferences for college degree programmes. *The Annals of Applied Statistics* 8 1145–1181.

Celeux, G., Hurn, M. & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* 95 957–970.

Chen, M.-H., Shao, Q.-M. & Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. New York: Springer.

Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data*, vol. 34 of *Lecture Notes in Statistics*. Springer-Verlag.

Critchlow, D. E., Fligner, M. A. & Verducci, J. S. (1991). Probability models on rankings. *J. Math. Psych.* 35 294–318.

Croon, M. A. (1989). Latent class models for the analysis of rankings. In G. De Soete, H. Feger & K. C. Klauer, eds., *New developments in psychological choice modeling*. Elsevier Science Publisher B. V., 99.

Croon, M. A. & Luijkx, R. (1993). Latent structure models for ranking data. In *Probability models and statistical analyses for ranking data (Amherst, MA, 1990)*, vol. 80 of *Lecture Notes in Statist.* New York: Springer, 53–74.

Daniels, H. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 12 171–191.

David, H. A. (1988). *The method of paired comparisons*, vol. 41 of *Griffin's Statistical Monographs & Courses*. Charles Griffin & Co. Ltd., 2nd ed.

Dayton, C. M. & Macready, G. B. (1988). Concomitant-variable latent-class models. *Journal of the American Statistical Association* 83 173–178.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 39 1–38. With discussion.

Diaconis, P. W. (1988). *Group representations in probability and statistics*, vol. 11 of *IMS Lecture Notes Monogr. Ser.* Inst. of Math. Stat.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 363–375.

Feigin, P. D. & Cohen, A. (1978). On a model for concordance between judges. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 40 203–213.

Fligner, M. A. & Verducci, J. S. (1986). Distance based ranking models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 48 359–369.

Fligner, M. A. & Verducci, J. S. (1988). Multistage ranking models. *J. Amer. Statist. Assoc.* 83 892–901.

Fraley, C. & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *J. Classification* 20 263–286.

Gabrielli, F., Salvi, R., Garulli, C., Kalogris, C., Arima, S., Tardella, L., Monaci, P., Pupa, S. M., Tagliabue, E., Montani, M., Quaglino, E., Curcio, C., Marchini, C. & Amici, A. (2013). Identification of relevant conformational epitopes on the HER2 oncoprotein by using Large Fragment Phage Display (LFPD). *PlosONE* 8.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2nd ed.

Gormley, I. C. & Murphy, T. B. (2006). Analysis of irish third-level college applications data. *Journal of the Royal Statistical Society: Series A* 169 361–379.

Gormley, I. C. & Murphy, T. B. (2008a). Exploring voting blocs within the irish electorate: a mixture modeling approach. *J. Amer. Statist. Assoc.* 103 1014–1027.

Gormley, I. C. & Murphy, T. B. (2008b). A mixture of experts model for rank data with applications in election studies. *Ann. Appl. Stat.* 2 1452–1477.

Gormley, I. C. & Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Anal.* 4 265–295.

Gormley, I. C. & Murphy, T. B. (2010). Clustering ranked preference data using sociodemographic covariates. In S. Hess & A. Daly, eds., *Choice Modelling: The State-of-the-Art and the State-of-Practice: Proceedings from the Inaugural International Choice Modelling Conference*. Emerald, 543–569.

Grün, B. & Leisch, F. (2009). Dealing with label switching in mixture models under genuine multimodality. *Journal of Multivariate Analysis* 100 851–861.

Guiver, J. & Snelson, E. (2009). Bayesian inference for Plackett-Luce ranking models. In L. Bottou & M. Littman, eds., *Proceedings of the 26th International Conference on Machine Learning – ICML 2009*. Omnipress, 377–384.

Gupta, J. & Damien, P. (2002). Conjugacy class prior distributions on metric-based ranking models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology* 64 433–445.

Gustafson, P. (2010). Bayesian inference for partially identified models. *The international journal of biostatistics* 6.

Henery, R. J. (1983). Permutation probabilities for Gamma random variables. *J. Appl. Probab.* 20 822–834.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *Ann. Statist.* 32 384–406.

Hunter, D. R. & Lange, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* 58 30–37.

Hurn, M., Justel, A. & Robert, C. P. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12 55–79.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation* 3 79–87.

Jasra, A., Holmes, C. & Stephens, D. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* 50–67.

Johnson, T. R. & Kuhn, K. M. (2013). Bayesian Thurstonian models for ranking data using JAGS. *Behavior research methods* 45 857–872.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning* 37 183–233.

Knuth, D. E. (2005). *The art of computer programming*. Pearson Education.

Koopmans, T. C. & Reiersol, O. (1950). The identification of structural characteristics. *The Annals of Mathematical Statistics* 165–181.

Lange, K., Hunter, D. R. & Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graph. Statist.* 9 1–59. With discussion, and a rejoinder by Hunter and Lange.

Lebanon, G. & Mao, Y. (2008). Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.* 9 2401–2429.

Lee, P. H. & Yu, P. L. H. (2010). Distance-based tree models for ranking data. *Comput. Statist. Data Anal.* 54 1672–1682.

Lee, P. H. & Yu, P. L. H. (2012). Mixtures of weighted distance-based models for ranking data with applications in political studies. *Comput. Stat. Data Anal.* 56 2486–2500.

Lo, A. Y. et al. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The annals of statistics* 12 351–357.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*, vol. 7. Cambridge university press.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika* 44 114–130.

Marden, J. I. (1995). *Analyzing and modeling rank data*, vol. 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall.

Marin, J.-M., Mengersen, K. & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics* 25 459–507.

Marin, J.-M. & Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer.

McLachlan, G. & Krishnan, T. (2007). *The EM algorithm and extensions*, vol. 382. John Wiley & Sons.

McLachlan, G. & Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McNicholas, P. D., Murphy, T. B., McDaid, A. F. & Frost, D. (2010). Serial and parallel implementations of model-based clustering *via* parsimonious gaussian mixture models. *Comput. Statist. Data Anal.* 54 711–723.

Meilă, M. & Bao, L. (2008). Estimation and clustering with infinite rankings. In D. McAllester & P. Myllymaki, eds., *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence – UAI 2008*. AUAI Press, 393–402.

Meilă, M. & Bao, L. (2010). An exponential model for infinite rankings. *J. Mach. Learn. Res.* 11 3481–3518.

Meilă, M. & Chen, H. (2010). Dirichlet process mixtures of Generalized Mallows models. In P. D. Grünwald & P. Spirtes, eds., *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence – UAI 2010*. AUAI Press, 358–367.

Minka, T. (2004). Power EP. Tech. rep., Microsoft Research, Cambridge.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference*. Ph.D. thesis, Massachusetts Institute of Technology.

Mollica, C. & Tardella, L. (2014). Epitope profiling via mixture modeling of ranked data. *Statistics in Medicine* 33 3738–3758.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika* 16 3–9.

Murphy, T. B. & Martin, D. (2003). Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.* 41 645–655.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Statist.* 9 249–265.

Papastamoulis, P. & Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in Bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics* 19.

Paulino, C. D. M. & de Bragança Pereira, C. A. (1994). On identifiability of parametric statistical models. *Journal of the Italian Statistical Society* 3 125–151.

Philip, L. (2000). Bayesian analysis of order-statistics models for ranking data. *Psychometrika* 65 281–299.

Plackett, R. L. (1975). The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.* 24 193–202.

Plummer, M. et al. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003). March.* 20–22.

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

Richardson, S. & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59 731–792.

Robert, C. P. & Casella, G. (2004). *Monte Carlo statistical methods*, vol. 319. Springer.

Rodríguez, C. E. & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics* 23 25–45.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica: Journal of the Econometric Society* 577–591.

Rubin, D. B. et al. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics* 12 1151–1172.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate Bayesian inference for latent gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 319–392.

Savage, L. J. (1962). *The foundations of statistical inference.* A discussion opened by Professor L. J. Savage at a meeting of the JointStatistics Seminar, Birkbeck and Imperial Colleges, in the University of London. Methuen & Co., Ltd., London; John Wiley & Sons, Inc., New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6 461–464.

Shieh, G. S. (1998). A weighted Kendall's tau statistic. *Statist. Probab. Lett.* 39 17–24.

Shieh, G. S., Bai, Z. & Tsai, W.-Y. (2000). Rank tests for independence-with a weighted contamination alternative. *Statistica Sinica* 10 577–594.

Silverberg, A. R. (1980). *Statistical models for q-permutations.* ProQuest LLC, Ann Arbor, MI. Thesis (Ph.D.)–Princeton University.

Sperrin, M., Jaki, T. & Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing* 20 357–366.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 583–639.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62 795–809.

Tarsitano, A. (2009). Comparing the effectiveness of rank correlation statistics. *Working paper* 6.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics* 145 505–518.

Thompson, G. L. (1993). Generalized permutation polytopes and exploratory graphical methods for ranked data. *Ann. Statist.* 21 1401–1430.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Reviews* 34 273–286.

Train, K. E. (2003). *Discrete choice methods with simulation.* Cambridge University Press.

Vaida, F. (2005). Parameter convergence for EM and MM algorithms. *Statistica Sinica* 15 831.

Yao, G. & Böckenholt, U. (1999). Bayesian estimation of Thurstonian ranking models based on the Gibbs sampler. *British Journal of Mathematical and Statistical Psychology* 52 79–92.