A generative model for rank data based on an insertion sorting algorithm

Christophe Biernacki, Julien Jacques

A generative model for rank data based on an insertion sorting algorithm

Christophe BIERNACKI [a], Julien JACQUES [a]

Abstract

An original and meaningful probabilistic generative model for full rank data modelling is proposed. Rank data arise from a sorting mechanism which is generally unobservable for statisticians. Assuming that this process relies on paired comparisons, the insertion sort algorithm is known as being the best candidate in order to minimize the number of potential paired misclassifications for a moderate number of objects to be ordered. Combining this optimality argument with a Bernoulli event during paired comparison step, a model that possesses desirable theoretical properties, among which are unimodality, symmetry and identifiability is obtained. Maximum likelihood estimation can also be performed easily through an EM or a SEM-Gibbs algorithm (depending on the number of objects to be ordered) by involving the latent initial presentation order of the objects. Finally, the practical relevance of the proposal is illustrated through its adequacy with several real data sets and a comparison with a standard rank data model.

Key words and phrases. Full rank data, sorting process, insertion sort algorithm, EM algorithm, quiz data.

1 Introduction

Ranking data is of great interest in human activities involving preferences, attitudes or choices like Web Page ranking, Sport, Politics, Economics, Educational Testing, Biology, Psychology, Sociology, Marketing, *etc*. Ranks are so meaningful that it is not unusual for them to be found as the result of a transformation of other kinds of data. In this paper only full rankings will be considered but possible extensions for partial, tied or incomplete rankings are being considered as future prospects of the proposed model.

The rank datum results from a ranking of m objects $\mathscr{O}_1,\ldots,\mathscr{O}_m$ by a judge (human or not). Two representations of these data are commonly used: ranking and ordering. The *ranking* representation $x^{-1}=(x_1^{-1},\ldots,x_m^{-1})$ contains the ranks assigned to the objects, and means that $\mathscr{O}_i$ is in x_i^{-1}th position ($i=1,\ldots,m$). A ranking is then an element of $\mathscr{P}_m$,

[a] Laboratoire P. Painlevé, UMR 8524 CNRS Université Lille I, Bât M2, Cité Scientifique, F-59655 Villeneuve d'Ascq Cedex, France.

the set of permutations of the first m integers. The *ordering* representation $x = (x_1, \ldots, x_m)$ is also an element of $\mathscr{P}_m$ and means that Object $\mathscr{O}_{x_i}$ is in the ith position ($i = 1, \ldots, m$). Let us consider the following example to illustrate both notations: a judge, which has to rank three holidays destinations according to its preferences, $\mathscr{O}_1 =$ Countryside, $\mathscr{O}_2 =$ Mountain and $\mathscr{O}_3 =$ Sea, ranks first Sea, second Countryside, and last Mountain. The ordering result of the judge is $x = (3,1,2)$ whereas the ranking result is $x^{-1} = (2,3,1)$. In the sequel both ordering and ranking notations will be used for rank data.

Rank data are multivariate but highly structured data. So, beyond standard but general data analysis methods (means, factor analysis, *etc.*), some specific descriptive methods that respect this structure have been proposed. For instance the permutation polytope to plot the rank vectors in the Euclidean space (see [30, 31] and an example in Figure 1) and suitable distances to define the centre and spread of a data set [18, 26, 10].

From an inference point of view, distances are useful to test the distribution of these data (uniformity, populations comparison, see [9, 27]) or to model the distribution itself (for instance the Mallows Φ model relies on the Kendall distance, see [18, 6]). More generally, parametric probabilistic models, if relevant and allowing easy parameter interpretation, are useful to summarize and understand such complex data and are basic tool for density estimation, prediction or clustering. Major rank data models date from the mid 20th century and most of the current works on the topic use these models.

The two most popular classes of models for rank data consist on modelling directly the hypothetical ranking process followed by the judge. For a complete review, refer to [27, Chap. 5 to 10]. The first class is derived from a paired comparison process [19]: the judge constructs a rank by first comparing each pair of objects, and second ensuring the consistency of these paired comparisons (if $\mathscr{O}_1$ is prefered to $\mathscr{O}_2$ and $\mathscr{O}_2$ to $\mathscr{O}_3$, $\mathscr{O}_1$ must be prefered to $\mathscr{O}_3$). It leads to the *Babington Smith model* [27, p. 116]. The number of parameters of this model is very large especially when m grows, therefore some simplifications have been considered. [4] associate to each object $\mathscr{O}_j$ a score indicating an overall degree of preference of this object, and connect these scores to the Babington Smith model's parameters, which defines the *Bradley–Terry–Mallows model* [27, p. 117]. [26] goes forward into the simplification by imposing that the scores only depend on the sign of $\mu_{x_i}^{-1} - \mu_{x_j}^{-1}$, where μ is a "reference" rank. It leads to the famous *Mallows Φ model*:

$$\mathrm{p}(x;\mu,\lambda) = \mathscr{C}(\lambda)^{-1} \exp^{-\lambda K(x,\mu)},$$

where K is the Kendall distance between two ranks [18, 6] and where

$$\mathscr{C}(\lambda) = \prod_{j=1}^{m-1} \frac{1-\exp(-(m-j+1)\lambda)}{1-\exp(-\lambda)}$$

is a normalization constant [10] with $\lambda \in \mathbb{R}$ a precision parameter. For instance, a high λ value leads to strong unimodality around μ. In the last decade, some extensions of this model have been proposed: [22] use a generalization of Mallows Φ model to combine multiple input rankings, and [23] model partially ranked data *via* semi-parametric model using Mallows kernel.

The second popular class of rank data models is multistage models, which consider the following iterative ranking process: the judge selects firstly the best object among the m ones, then the best among the $m-1$ remaining ones, and so on. The *Plackett-Luce model* [25, 29] is then based on the probability that each $\mathscr{O}_{x_i}$ is ranked first among the m objects [27, p. 119]. It could be noticed that this model corresponds to Thurstonian model [32, 3] with a Gumbel density. [1] proposed a variant by introducing a dampening parameter and [14] combined the Plackett-Luce model with a latent space model in order to spatially model the ranked nature of the data. [11] introduce an alternative multistage model by considering another form of probability at each step of the ranking process. It leads to the Fligner and Verducci's *strongly unimodal model* [27, p. 120]. Assuming specific forms of this model could lead to Mallows Φ model or to a generalization of this latter named Φ component-model [27, p. 121].

Thus, the ranking processes that have motivated these two classes of rank data models can be interpreted as two different sorting processes, in which stochastic errors are introduced to define a probability distribution on the whole rank data space. The natural question involved by this interpretation is whether the used sorting algorithms are the most appropriate. For instance, in paired comparison models, it is not optimal to do so many comparisons since it leads to a sorting algorithm with excessively high computational complexity. Here, we propose a generative model for rank data based on (straight) *insertion* sort algorithm. This is one of the most powerful sorting algorithms among the usual ones when $m \leq 10$ [21, Chap. 5], the situation expected to be the most frequent particularly in case of "human ranking". However, the proposed model can be applied in practice to any number of objects. This new kind of generative model enjoys good theoretical properties and has originality to involve the (potentially latent) initial presentation order of the objects.

The paper is organized as followed. Section 2 introduces and sets up the proposed model which is based on insertion sort algorithm, and its theoretical properties (unimodality, symmetry, identifiability) are also detailed. Maximum likelihood estimation is considered in Section 3 by means of an EM or a SEM-Gibbs algorithm (depending on the number of objects to be ordered) since a missing data interpretation of the proposed model can be exhibited. Numerical illustrations are presented in Section 4 to evaluate the relevance of the proposed model on real data sets both from a distributional adequacy and comparison to the usual Mallows Φ model point of view. Since this work sheds a new light on rank data modelling, numerous related perspectives are discussed in the last section (Section 5).

2 A generative model for rank data based on insertion sort

2.1 Motivation to adopt insertion sort algorithm

We assume it exists an ordering $\mu = (\mu_1, \ldots, \mu_m)$ on the m objects, so that a judge who perfectly sorts these objects returns this *reference* rank μ. Making also the natural as-

Table 1: An example to illustrate the standard insertion sort process with $\mu = (1,2,3)$ and $y = (1,3,2)$

step j	unsorted	sorted
start	$y = \boxed{1}\boxed{3}\boxed{2}$	-
1	$\boxed{3}\boxed{2}$	$\boxed{1}$
2	$\boxed{2}$	$\boxed{3} \overset{?}{\leftrightarrow} \boxed{1}$ $\boxed{1}\boxed{3}$
3	-	$\boxed{2} \overset{?}{\leftrightarrow} \boxed{1}\boxed{3}$ $\boxed{1}\boxed{2} \overset{?}{\leftrightarrow} \boxed{3}$ $x = \boxed{1}\boxed{2}\boxed{3}$

sumption that a rank $x = (x_1, \ldots, x_m)$ is the result of a sorting process relying on successive object paired comparisons, any difference between the final rank x and μ is necessarily attributed to some incorrect paired comparisons. As a consequence, reducing the gap between x and μ is strongly correlated to minimize the number of paired comparisons involved in the sorting process. For instance, for a moderate number of objects ($m \leq 10$), an "optimal judge" should adopt the *insertion* sort algorithm which is optimal in this case [21, Chap. 5]. Arguing that $m \leq 10$ is a frequent situation in practice (in particular for "human rankings"), we propose to keep this sorting algorithm although in practice our model will not impose any upper bound on m.

We briefly sketch how the standard insertion sort algorithm performs. More details can be found in [21, Chap. 5] for instance. We note $y = (y_1, \ldots, y_m)$ the initial order of the objects presented to the judge. First, the current object in y to be sorted is placed before the already sorted objects, and is compared to the first object after. If the relative position of both objects in this pair is correct (according to μ), this pair order is unchanged and this process is restarted with the next object in y. Otherwise, the pair order is reversed and a new pair comparison is performed with the next object after (if it exists). And so forth, until obtaining the final ranking x. An example is also displayed in Table 1 with $m = 3$ to give step by step an overview of standard insertion sort algorithm.

2.2 Modelling a stochastic insertion sort algorithm

Our idea is to merge the previous deterministic process with the following stochastic paired comparisons process. We assume that each paired comparison is the result of a Bernoulli experiment whose outcome is a correct comparison (according to μ) with probability $\pi \in [0,1]$ and an incorrect comparison with probability $1-\pi$. The parameter π models the reliability of the judge about the "true" rank μ. We also assume that each pair ranking operation is independent from others and that the probability π is constant along the sorting process. Thus, at the end of the proposed stochastic process, the final rank x can differ from the reference rank μ. From these simple ideas, we obtain a meaningful generative model for rank data that is presented at length now.

Based on this stochastic modelling of the insertion sort algorithm, the probability

$\mathrm{p}(x|y;\mu,\pi)$ to obtain a rank x from an initial presentation order y, with a reference rank μ, is given by:

$$\mathrm{p}(x|y;\mu,\pi) = \pi^{G(x,y,\mu)}(1-\pi)^{A(x,y)-G(x,y,\mu)}, \tag{2.1}$$

where π is the probability of *good* paired comparison (according to μ), $G(x,y,\mu)$ is the number of good paired comparisons in the sorting process, and $A(x,y)$ is the overall number of paired comparisons. In addition, denoting by $\mathrm{p}(y)$ the presentation order distribution, the marginal distribution of x is given by:

$$\mathrm{p}(x;\mu,\pi) = \sum_{y\in\mathscr{P}_m} \mathrm{p}(x|y;\mu,\pi)\mathrm{p}(y). \tag{2.2}$$

In this paper, we assume the presentation orders are unknown and uniformly distributed, thus $\mathrm{p}(y) = m!^{-1}$ for all $y \in \mathscr{P}_m$. In the sequel the rank data model defined by distribution (2.2) will be quoted as ISR(μ,π) for Insertion Sort Rank data model associated to parameters (μ,π).

The proof of (2.1) is given in B. At the beginning of this appendix, notations $A(x,y)$ and $G(x,y,\mu)$ are also mathematically defined and illustrated through an example in Table 5. Note that Tables 1 (deterministic insertion algorithm) and 5 (stochastic insertion algorithm) are different because they lead to a different x value.

Remark The fact that π remains constant along the sorting process makes sense as the judge's knowledge does not change on μ as well as the tiredness of the judge is negligible. However, this hypothesis could be weakened. This issue is discussed in Section 5.

2.3 Properties of ISR model

In this section, the main properties of the ISR model are stated. Precise statement of each property as a mathematical proposition and related proofs are available in C at the end of this paper. Proofs rely on applying permutation properties on both ranking and ordering notations on $\mathscr{P}_m$.

- *Uniformity for $\pi = \frac{1}{2}$.* In the case where paired comparison is performed at random ($\pi = \frac{1}{2}$), the ISR is the uniform distribution on $\mathscr{P}_m$. In this case, the reference rank μ can be arbitrarily chosen. See Proposition 3 in C.

- *Mode μ.* One of the most important properties which can be expected from the ISR distribution is that the reference rank μ is the unique mode of the distribution if $\pi > \frac{1}{2}$. See Proposition 4 in C.

- *Anti-mode $\bar{\mu}$.* Let $\bar{\mu}$ be defined by $\bar{\mu} = \mu \circ \bar{e}$ where $\bar{e} = (m,\ldots,1)$ is the permutation of total inversion. This rank $\bar{\mu}$ is the farthest from μ for the Kendall distance. Then the unique anti-mode (the rank of smallest probability) is $\bar{\mu}$ if $\pi > \frac{1}{2}$. See Corollary 1 in C.

- *Link between μ and π.* The mode μ is also uniformly more pronounced when π grows. See Proposition 5 in C.

- *Symmetry.* Distributions $\text{ISR}(\mu,\pi)$ and $\text{ISR}(\overline{\mu},1-\pi)$ are equivalent. This property will be especially useful to exhibit the identifiability conditions of the ISR distribution below. See Proposition 6 in C.
- *Identifiability.* The uniformity for $\pi=\frac{1}{2}$ of the ISR distribution and its symmetry lead to impose $\pi>\frac{1}{2}$ as a necessary condition for identifiability. In fact this condition is also sufficient to have identifiability. See Proposition 7 in C.

3 Estimation of the model parameters

The ISR model for rank data has two parameters: the probability $\pi\in[\frac{1}{2},1]$ and the reference rank, or modal rank, μ, which can take its values in $\mathscr{P}_m$. Note that the case $\pi=\frac{1}{2}$ is kept, although this is a non-identifiability situation because it leads to the uniformity of the ISR distribution, that can be of interest for practical applications. Considering $(x^1,\ldots,x^n)$ as an independent sample of n ranks from $\text{ISR}(\mu,\pi)$, we present in this section estimation of (μ,π) by maximizing the log-likelihood of the ISR model given by

$$l(\mu,\pi;x^1,\ldots,x^n)=\sum_{i=1}^{n}\ln\left(\frac{1}{m!}\sum_{y\in\mathscr{P}_m}\mathrm{p}(x^i|y;\mu,\pi)\right).$$

We assume in the following that pairs (x^i,y^i) arise independently ($i=1,\ldots,n$), where $(y^1,\ldots,y^n)$ denote all the latent presentation orders.

3.1 Using an EM algorithm for a small number of objects ($m\leq 7$)

As $(y^1,\ldots,y^n)$ are unknown, we use an EM algorithm [7] to maximize the *observed* data log-likelihood. Denoting by $(\mu,\pi)^{\{0\}}$, the parameter starting values for the EM algorithm and by $(\mu,\pi)^{\{q\}}$ the value of the parameters at the step q ($q\in\mathbb{N}$), the two steps (E and M) of this algorithm are described as follows. This algorithm is computationally feasible for a small number of objects, typically $m\leq 7$. However, it could lead to computational difficulties for larger m values because of the sum over $\mathscr{P}_m$ involved in both E and M steps (see discussion at the beginning of Section 3.3).

Note that since the conditional probability (2.1) is invariant to an inversion of the first two elements of the presentation order (Lemma 1 of D), the number $m!$ of presentation orders y to be considered in the calculation of the probability (2.2) may be reduced by half. This remark can be used in order to accelerate the EM algorithm.

The E step The *complete-data* log-likelihood is given by

$$l_c(\mu,\pi;x^1,\ldots,x^n,y^1,\ldots,y^n)=\sum_{i=1}^{n}\sum_{y\in\mathscr{P}_m}\mathbf{1}\{y=y^i\}\ln\left(\frac{1}{m!}\mathrm{p}(x^i|y;\mu,\pi)\right).$$

The E step consists of computing the conditional expectation $\mathscr{Q}$ of l_c expressed by:

$$\mathscr{Q}((\mu,\pi),(\mu,\pi)^{\{q\}};x^1,\ldots,x^n)=\sum_{i=1}^{n}\sum_{y\in\mathscr{P}_m}t_{iy}^{\{q\}}\ln\left(\frac{1}{m!}\mathrm{p}(x^i|y;\mu,\pi)\right),$$

where the conditional probability $t_{iy}^{\{q\}}$ that $y^i = y$ is given by:

$$t_{iy}^{\{q\}} = \frac{\mathrm{p}(x^i|y;(\mu,\pi)^{\{q\}})}{\sum_{\tau\in\mathscr{P}_m}\mathrm{p}(x^i|\tau;(\mu,\pi)^{\{q\}})}.$$

The M step The M step consists of choosing the value $(\mu,\pi)^{\{q+1\}}$ which maximizes the conditional expectation $\mathscr{Q}$ computed at the E step:

$$(\mu,\pi)^{\{q+1\}} = \underset{(\mu,\pi)\in\mathscr{P}_m\times[\frac{1}{2},1]}{\operatorname{argmax}} \mathscr{Q}((\mu,\pi),(\mu,\pi)^{\{q\}};x^1,\ldots,x^n).$$

For the modal rank μ, it is however numerically expensive to explore all the discrete space $\mathscr{P}_m$ even for relatively small values of m. To overcome this difficulty, a specific strategy will be proposed in Section 3.4. The value of probability π, maximizing $\mathscr{Q}$ leads to following update:

$$\pi^{\{q+1\}} = \frac{\sum_{i=1}^n\sum_{y\in\mathscr{P}_m} t_{iy}^{\{q\}} G(x^i,y,\mu^{\{q\}})}{\sum_{i=1}^n\sum_{y\in\mathscr{P}_m} t_{iy}^{\{q\}} A(x^i,y)}.$$

Note that this value of $\pi^{\{q+1\}}$ can be interpreted as the proportion of good manipulations (switching to right or stop) in the insertion sort algorithm.

The algorithm stops when the difference of the log-likelihood between two successive iterations is less than a given threshold. We discuss now how to efficiently start the algorithm.

3.2 Initialization strategy for π in EM

We propose here a straightforward asymptotic bound on π to initialize EM. We will discuss in Section 3.4 how to also restrict the possible values for μ.

Proposition 1. *Denoting by f_0 the empirical modal relative frequency, the interval $[\hat{\pi}^-,\hat{\pi}^+]$ asymptotically contains π where*

$$\hat{\pi}^- = f_0^{\frac{1}{m-1}} \text{ and } \hat{\pi}^+ = f_0^{\frac{2}{m(m-1)}}. \tag{3.1}$$

Proof. Using Lemma 6 (see D) and also the fact that, for any μ and y, $\mathrm{p}(\mu|y;\mu,\pi) = \pi^{A(\mu,y)}$ (see the proof in Lemma 5), it leads to the following bounds for the probability of μ:

$$\pi^{m(m-1)/2} \leq \mathrm{p}(\mu;\mu,\pi) \leq \pi^{m-1}.$$

Since f_0 is a consistent estimator of $\mathrm{p}(\mu;\mu,\pi)$, it ends the proof. □

As soon as $\hat{\pi}^-$ and $\hat{\pi}^+$ are greater than $\frac{1}{2}$, this result is useful to initialize π in EM by choosing uniformly $\pi^{\{0\}}$ in the interval given by (3.1). If only $\hat{\pi}^+ \geq \frac{1}{2}$, the interval becomes $[\frac{1}{2},\hat{\pi}^+]$. If both bounds are lower than $\frac{1}{2}$, then the interval $[\frac{1}{2},1]$ must be used. In Table 4 of Section 4, bounds associated with all the data sets are greater than $\frac{1}{2}$ and the retained intervals are quite narrow in comparison to $[\frac{1}{2},1]$, so the strategy seems to be efficient.

3.3 Using a SEM-Gibbs algorithm for a large number of objects ($m \geq 8$)

As discussed at the beginning of Section 3.1, the EM algorithm is not appropriate as soon as $m \geq 8$ because of the factorial term present in both E and M steps: for $m = 8$ or $m = 9$, EM could be still feasible although it would be extremely slow (respectively about 4.10^4 and 4.10^5 sums involve in E and M steps); for $m \geq 10$, this computational times would become definitively inaccessible (more than 3.10^6 sums involve in E and M steps).

A so-called SEM-Gibbs algorithm may provide an efficient and elegant solution. The fundamental idea of this algorithm is to reduce the computational complexity that was present in both E and M steps of EM by removing all explicit and extensive use of the conditional probabilities $t_{iy}^{\{q\}}$. It relies on the SEM algorithm [12, 5] which generates the latent variables y^i at a so-called stochastic step (S step) from the conditional probabilities $t_{iy}^{\{q\}}$ computed at the E step. Then these latent variables are directly used during the M step. However, the advantage with SEM-Gibbs algorithm relies on the fact that the latent variables are generated without calculating conditional probabilities thanks to Gibbs algorithm. The proposed SEM-Gibbs algorithm proceeds in the following steps:

The SE-Gibbs step It consists of generating a sample $y^{i\{q\}}$ from $t_{iy}^{\{q\}}$ (for all $i \in \{1,\ldots,n\}$) like in a SE step of the standard SEM algorithm but without any calculation of conditional probabilities $t_{iy}^{\{q\}}$ since it invokes instead the following Gibbs algorithm. Starting from an arbitrary sample $y^{i\{q,0\}}$, generate $r \in \{1,\ldots,R\}$ sequences $y^{i\{q,r\}}$ (R being a given number) where

$$(y_j^{i\{q,r+1\}},\cdot) \sim p\left(y_j^i, y_{j+1}^i | (y_1,\ldots,y_{j-1})^{i\{q,r+1\}}, (y_{j+2},\ldots,y_m)^{i\{q,r\}}, x^1,\ldots,x^n; (\mu,\pi)^{\{q\}}\right)$$

for $j \in \{1,\ldots,m-2\}$ and where

$$(y_{m-1},y_m)^{i\{q,r+1\}} \sim p\left(y_{m-1}^i, y_m^i | (y_1,\ldots,y_{m-2})^{i\{q,r+1\}}, x^1,\ldots,x^n; (\mu,\pi)^{\{q\}}\right).$$

Note that both previous expressions do not involve any combinatorial calculation. If R is large enough, $y^{i\{q,R\}}$ arises from $t_{iy}^{\{q\}}$, thus we retain $y^{i\{q\}} = y^{i\{q,R\}}$ ($i = 1,\ldots,n$).

The M step The M step consists of choosing the value $(\mu,p)^{\{q+1\}}$ which maximizes the completed log-likelihood computed at the SE-Gibbs step:

$$(\mu,\pi)^{\{q+1\}} = \underset{(\mu,\pi)\in\mathscr{P}_m\times[\frac{1}{2},1]}{\operatorname{argmax}} \; l_c(\mu,\pi;x^1,\ldots,x^n,y^{1\{q\}},\ldots,y^{n\{q\}}).$$

For the modal rank μ, it is however numerically expensive to explore all the discrete space $\mathscr{P}_m$ even for relatively small values of m. To overcome this difficulty, a specific strategy will be proposed in Section 3.4. For the probability π, maximizing $l_c(\mu,\pi;x^1,\ldots,x^n,y^{1\{q\}},\ldots,y^{n\{q\}})$ leads to the following update, removing any combinatorial difficulty:

$$\pi^{\{q+1\}} = \frac{\sum_{i=1}^n G(x^i,y^{i\{q\}},\mu^{\{q\}})}{\sum_{i=1}^n A(x^i,y^{i\{q\}})}.$$

Unlike EM, the random sequence of parameters $(\mu,\pi)^{\{q\}}$ generated by SEM-Gibbs does not converge pointwise. Consequently its stopping rule can not rely on difference of the likelihood between two successive iterations. The simplest alternative solution is to stop the SEM-Gibbs algorithm after a given number Q of iterations. After removing a burn in period corresponding to the first B iterations, we retain a point estimate of (μ,π) in the following manner: for each distinct μ values in the sequence $\mu^{\{q\}}$ $(q=B,\ldots,Q)$, take the mean $\overline{\pi}_\mu$ of the associated $\pi^{\{q\}}$ values and then retain the couple $(\mu,\overline{\pi}_\mu)$ leading to the highest log-likelihood $l(\mu,\overline{\pi}_\mu;x^1,\ldots,x^n)$. Since the log-likelihood calculation suffers from combinatorial issues, we use in addition the following approximation:

$$l(\mu,\pi;x^1,\ldots,x^n)=-\sum_{i=1}^{n}\ln\left(\sum_{y\in\mathscr{P}_m}\frac{1}{p(x^i|y;\mu,\pi)}p(y|x^i;\mu,\pi)\right)\approx-\sum_{i=1}^{n}\ln\left(\frac{1}{S}\sum_{s=1}^{S}\frac{1}{p(x^i|y^{i,s};\mu,\pi)}\right),$$

where all $y^{i,s}$ arise independently from $p(y|x^i;\mu,\pi)$ $(s=1,\ldots,S)$. Section 3.5 will validate on some real and artificial data sets that our SEM-Gibbs strategy leads to very accurate estimates despite its lower computational cost compared to the standard EM algorithm.

3.4 Reducing the number of reference ranks μ for EM and SEM-Gibbs

We propose a strategy to reduce (often drastically) the number of possible values for μ in the step M of both EM and SEM-Gibbs. This result relies on Proposition 1 and also on the following proposition.

Proposition 2. *Let N_x be the number of individuals equal to $x\in\mathscr{P}_m$ among a random sample from* $\text{ISR}(\mu,\pi)$ *of size n. Denoting by*

$$h_\alpha(\pi)=\#\{x:p(N_x\geq N_\mu;\mu,\pi)\geq\alpha\}$$

the number of ranks for which the empirical frequency can be greater than or equal (with probability at least $\alpha\in[0,1]$) to the empirical frequency associated with the theoretical modal rank μ, the following inequality asymptotically holds for any $\mu\in\mathscr{P}_m$ and $\pi\in[\frac{1}{2},1]$:

$$h_\alpha(\pi)\leq h_\alpha\left(\hat{\pi}^-\right).$$

Proof. We know from Proposition 1 that asymptotically $\hat{\pi}^-\leq\pi$. Proposition 5 in C allows to conclude the proof. □

The idea is to browse the empirical modal rank in association with some other ranks having high empirical relative frequency. The following strategy can be used only if $\hat{\pi}^-\geq\frac{1}{2}$. Firstly, $h_\alpha(\hat{\pi}^-)$ is estimated with a parametric bootstrap [8] of M replications from $\text{ISR}(\mu,\hat{\pi}^-)$. The key point is that it is independent from μ, so any $\mu\in\mathscr{P}_m$ can be used. Then the $h_\alpha(\hat{\pi}^-)$ most frequent distinct ranks in the sample $(x^1,\ldots,x^n)$ are retained as possible μ values among the potential $m!/2$ possibilities and are used both as potential initial values $\mu^{\{0\}}$ and also as values to browse during the M step.

The proposed strategy is aimed to significantly reduce the number of candidates for μ. The number of candidates for μ reduces when the size of the observed sample n grows since $h_\alpha(\hat{\pi}^-) \xrightarrow{p} 1$ when $n \to \infty$. So, the browsed ranks are asymptotically reduced to the empirical modal rank which is known to be a consistent estimator of μ.

Table 4 (Column "#μ") of Section 4 illustrates through numerical examples that this procedure effectively reduces the number of possible ranks for μ in comparison to the $m!/2$ possible values.

Remarks

- Proposition 2 gives asymptotic guaranties on the proposed strategy to reduce the number of candidates for mode μ of the distribution. The practical efficiency of this strategy is illustrated in the next section through simulation studies.
- On the real data sets used in Section 4, we noticed that the proposed strategy to preselect μ is less time-consuming since it takes less than 1% of the the entire process composed of preselection and of estimation (whatever be the estimation algorithm: EM or SEM-Gibbs).
- Note that the selection of the possible ranks should be carried out only once before the start of the EM or the SEM-Gibbs algorithms. This will be the strategy we follow in experiments throughout this paper.

3.5 Validation of the SEM-Gibbs algorithm based on simulations

In order to illustrate the estimation accuracy of the SEM-Gibbs approach, we propose to evaluate it both on simulated and real data sets. We introduce here the *normalized* Kendall distance $\bar{K}(\cdot,\cdot) = K(\cdot,\cdot)2[m(m-1)]^{-1}$ because its value is between 0 and 1, indicating respectively minimum and maximum disagreement.

- *Simulated data.* 20 samples of size $n = 100$ are simulated according to ISR with three values of π ($\pi \in \{0.6, 0.75, 0.9\}$) and two values of m ($m \in \{5, 10\}$). These parameters are then estimated by SEM-Gibbs approach with the following settings: $Q = 100$ SEM iterations, $B = 10$ burn in iterations and $R = 10$ Gibbs iterations inside the SE step. Results are displayed in Table 2. We tried with higher values for Q, B and R, but the results were essentially unchanged. Despite the small number of both Gibbs and SEM iterations, it appears that the accuracy of estimated parameters is very satisfactory, both with low and larger value of m. Note also that estimation of μ is a harder task when π is low, as expected (recall that information on μ is weak for small π).
- *Real data.* For all data sets considered below in Section 4 we have run EM and SEM-Gibbs, except the Election data set where $m = 14$ (EM is not numerically available in this case). Tuning parameters of EM and SEM-Gibbs are also the same as in Section 4. However, SEM-Gibbs is run here 10 times at random to evaluate stability of its results for μ, π and l (here l is exact likelihood, not the approximated

Table 2: Simulated data to validate SEM-Gibbs algorithm (20 replicates of size $n = 100$): $\bar{\hat{\pi}}$ and $\bar{K}(\hat{\mu}, \mu)$ indicate respectively the mean of the π estimate (in parenthesis its standard deviation) and the mean of the *standardized* Kendall distance between the μ estimate and the true one (in parenthesis its standard deviation).

	$m = 5$		$m = 10$	
π	$\bar{\hat{\pi}}$	$\bar{K}(\hat{\mu}, \mu)$	$\bar{\hat{\pi}}$	$\bar{K}(\hat{\mu}, \mu)$
0.6	0.572 (0.032)	0.210 (0.141)	0.566 (0.250)	0.180 (0.063)
0.75	0.759 (0.021)	0.000 (0.000)	0.700 (0.022)	0.056 (0.036)
0.9	0.895 (0.015)	0.000 (0.000)	0.884 (0.008)	0.002 (0.007)

Table 3: Real data to validate SEM-Gibbs algorithm (10 independent runs of SEM-Gibbs): $\hat{\mu}_{\text{EM}}$, $\hat{\mu}_{\text{SEM-gibbs}}$, $\hat{\pi}_{\text{EM}}$, $\hat{\pi}_{\text{SEM-gibbs}}$, l_{EM}, $l_{\text{SEM-Gibbs}}$ denote respectively μ, π and l estimated by EM and by SEM-Gibbs. Column "mean" displays the mean of each statistics on μ, π and l given at top of the table, over 10 runs of SEM-Gibbs. Columns "best" and "worst" give the value of these statistics with parameters obtained respectively with the best and the worst likelihood over the 10 runs of SEM-Gibbs.

	$K(\hat{\mu}_{\text{EM}}, \hat{\mu}_{\text{SEM-gibbs}})$			$\lvert\hat{\pi}_{\text{EM}} - \hat{\pi}_{\text{SEM-gibbs}}\rvert$			$l_{\text{EM}} - l_{\text{SEM-Gibbs}}$		
data set	mean	best	worst	mean	best	worst	mean	best	worst
Football	0.00	0.00	0.00	0.004	0.001	0.007	0.02	0.00	0.04
Cinema	0.00	0.00	0.00	0.003	0.000	0.006	0.01	0.00	0.02
Rugby	0.05	0.00	0.17	0.007	0.000	0.013	0.35	0.00	1.15
Words	0.00	0.00	0.00	0.001	0.000	0.002	0.02	0.01	0.02
Sports	0.01	0.00	0.05	0.002	0.000	0.005	0.09	0.00	0.40

one since it is numerically available for $m \leq 7$). Results are displayed in Table 3: in column "best", the best SEM-Gibbs (according to l) and EM coincide; in columns "means" and "worst" SEM-Gibbs is very close to EM. Thus, it confirms the good behaviour of SEM-Gibbs we validated previously with simulated data.

4 Numerical illustration

4.1 Presentation of the six real data sets

The ISR distribution is now compared to Mallows Φ model on six real data sets: two general knowledge quizzes (the answers of the 40 students being questioned are in E), four nations rugby league rankings, Fligner and Verducci's words associations rankings [10], Louis Roussos's sports rankings [27] and 2002 General Irish Election data set [14]. Mallows Φ model has been chosen as the reference model since, in addition to being one of the model based on paired comparison which has been the most studied [26, 10, 11,

28, 23], it is also linked with some multistage models (see Section 1).

- *Football* quiz. This quiz consists of ranking four national football teams according to increasing number of wins in the football World Cup: $\mathscr{O}_1 =$ France, $\mathscr{O}_2 =$ Germany, $\mathscr{O}_3 =$ Brasil, $\mathscr{O}_4 =$ Italy. The correct answer is $\mu^* = (1,2,4,3)$.

- *Cinema* quiz. This quiz consists of ranking chronologically the following Quentin Tarantino movies: $\mathscr{O}_1 =$ Inglourious Basterds, $\mathscr{O}_2 =$ Pulp Fiction, $\mathscr{O}_3 =$ Reservoir Dogs, $\mathscr{O}_4 =$ Jackie Brown. The correct answer is $\mu^* = (3,2,4,1)$.

- *Rugby*. This data set is the result of four nations rugby league, from 1883 to 1909 (except years 1888 and 1889 because only three nations were in the tournament, and except years 1886, 1890, 1897, 1898 and 1906 due to tie), which includes $\mathscr{O}_1 =$ England, $\mathscr{O}_2 =$ Scotland, $\mathscr{O}_3 =$ Ireland and $\mathscr{O}_4 =$ Wales.

- *Words*. [10] examined the data collected under the auspices of the Graduate Record Examination Board. A sample of 98 college students were asked to rank five words according to strength of association (least to most associated) with the target word "Idea": $\mathscr{O}_1 =$ Thought, $\mathscr{O}_2 =$ Play, $\mathscr{O}_3 =$ Theory, $\mathscr{O}_4 =$ Dream and $\mathscr{O}_5 =$ Attention.

- *Sports*. This data set is due to Louis Roussos [27] who asked 130 students at the University of Illinois to rank seven sports according to their preference in participating: $\mathscr{O}_1 =$ Baseball, $\mathscr{O}_2 =$ Football, $\mathscr{O}_3 =$ Basketball, $\mathscr{O}_4 =$ Tennis, $\mathscr{O}_5 =$ Cycling, $\mathscr{O}_6 =$ Swimming, $\mathscr{O}_7 =$ Jogging.

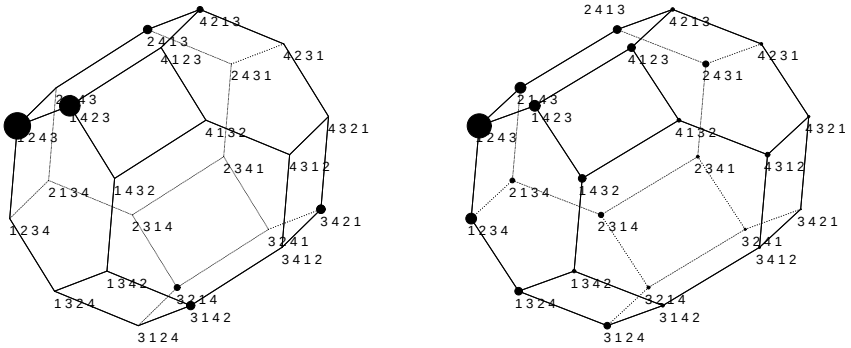
- *Election*. The general election system for the Irish House of Parliament is based on ranking of the candidates in order of preference. [14] presented and studied the 2002 General Election data set, consisting of the 64,081 rankings of the 14 candidates among which 2,490 are full rankings. Only these 2,490 full rankings are used here. Refer to [14] for more details on this data set.

The empirical distribution of the first three data sets (for which the number of objects to rank is 4) is graphically displayed on the *left* column of Figure 1 in the ranking space (orderings are displayed on each node).

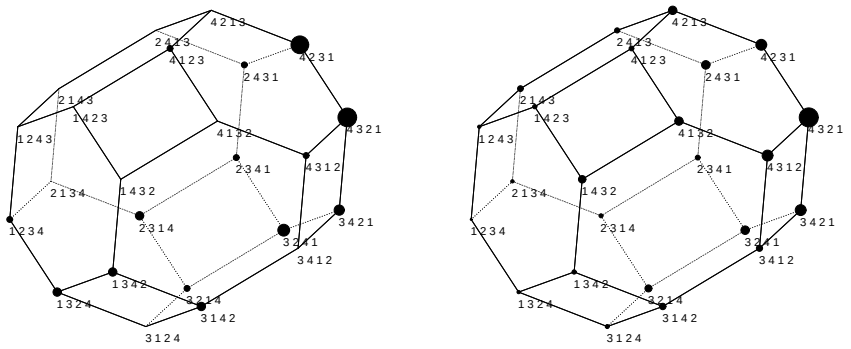
4.2 Estimation results

For each data set, the ISR distribution and the Mallows Φ model are estimated. An R package is available on the authors website[1]. For the ISR model, the estimation is carried out using EM algorithm when $m \leq 7$, and SEM-Gibbs algorithm is used when $m > 7$ (Election data set). For ISR the convergence threshold for the growth of log-likelihood in EM algorithm was fixed to 10^{-6} and only one initialization of π in $[\hat{\pi}^-, \hat{\pi}^+]$ has been used (no change on the results have been observed with several initializations), and $B = 30$, $Q = 100$, $R = 10$ for SEM-Gibbs algorithm is used. For Mallows Φ model, the numerical optimization has been carried out with a quasi-Newton method and the convergence threshold of ISR (10^{-6}).

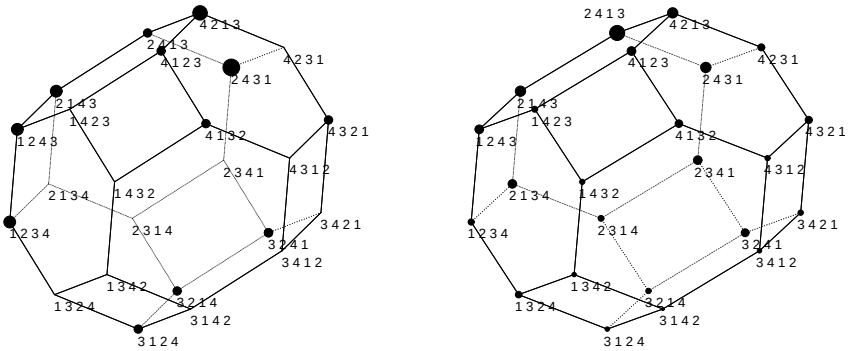
[1] `http://math.univ-lille1.fr/~jacques/soft.html`



(a) Football quiz data set



(b) Cinema quiz data set



(c) Rugby quiz data set

Figure 1: Empirical (left) and estimate ISR (right) distributions for the three data sets where $m = 4$. The area of the dots is proportional to the corresponding probability.

The ISR distribution of the first three data sets is graphically displayed on the *right* column of Figure 1 for a visual comparison with the empirical distribution. In addition, a χ^2 adequacy test, where the distribution under the null assumption is estimated by bootstrap [8] based on 1,000 replications, is performed for both models and for all data sets and the results are displayed in Table 4 (Column "$\widehat{\text{p-value}}$"). Depending on selected threshold on the p-value (typically models with p-value greater than 0.05 are not rejected) we notice that both models can be suitable for some data sets but not for all of them. Moreover, when comparing maximum log-likelihood values (Column "l"; the highest likelihoods are in bold), ISR leads to a greater maximum likelihood than Mallows Φ model for 4 data sets among 6. Consequently, ISR could be a natural competitor to be considered beside other classical models in any rank data analysis.

Table 4: ISR and Mallows Φ models estimation results: estimate parameters $\hat{\mu}$, $\hat{\pi}$ (ISR) and $\hat{\lambda}$ (Mallows), maximum log-likelihood l, estimated p-value of the χ^2 adequacy test, number of possible μ explored (#μ; For ISR it corresponds to $\hat{h}_\alpha(\hat{\pi}^-)$ with $\alpha = 0.05$ and $M = 100$ replications), lower and upper bounds $\hat{\pi}^-$ and $\hat{\pi}^+$ for π (ISR only)

data set	model	$\hat{\mu}$	$\hat{\pi}$ or $\hat{\lambda}$	l	$\widehat{\text{p-value}}$	#μ	$\hat{\pi}^-$	$\hat{\pi}^+$
Football	ISR	$(1,2,4,3)$	0.834	**-88.53**	0.001	1	0.794	0.891
	Mallows	$(1,2,4,3)$	1.106	-89.17	0.001	1	-	-
Cinema	ISR	$(4,3,2,1)$	0.723	**-111.94**	0.042	14	0.630	0.794
	Mallows	$(4,3,2,1)$	0.628	-112.12	0.029	2	-	-
Rugby	ISR	$(2,4,1,3)$	0.681	-58.68	0.538	12	0.585	0.765
	Mallows	$(2,4,1,3)$	0.528	**-58.33**	0.395	2	-	-
Words	ISR	$(2,5,4,3,1)$	0.879	-275.43	0.001	1	0.762	0.897
	Mallows	$(2,5,4,3,1)$	1.431	**-251.27**	0.019	1	-	-
Sports	ISR	$(1,3,2,4,5,7,6)$	0.564	**-1102.12**	0.999	$2^\dagger$	0.534	0.836
	Mallows	$(1,3,4,2,5,6,7)$	0.083	-1102.84	1	11	-	-
Election	ISR	$(13,4,1,2,3,5,6$ $7,8,9,10,11,12,14)$	0.682	**-48329.76**	0.999	6	‡	‡
	Mallows	$(4,13,2,5,1,14,7$ $6,10,8,9,12,3,11)$	0.164	-60157.38	0.999	38	-	-

† $\alpha = 0.1$ for the Sports data set to avoid to many μ due to the small π value ‡ Useless for SEM-Gibbs

We note that the strategy selecting the number of possible reference ranks to explore (Section 3.4) is effective. Indeed, only one candidate for μ has been selected by this strategy for the three data sets Football, Words and Sports (Column "#μ") and for other data sets the number of candidates is relatively small in comparison with $m!$. Concerning Mallows Φ model, the estimation of μ is carried out by an empirical iterative local search (in the sense of the Kendall distance) around the modal rank [11] which appears to be effective.

We discuss now the meaningful interpretation of ISR parameters by further analysing Table 4.

- *Football* quiz. The estimation of the reference rank μ coincides with the real rank. In addition, the accuracy level of students knowledge in football is quite high since

it is reflected by a high probability π (0.834) of well paired comparison. This underlines that the right answer is, on the whole, known with high level of confidence by this population of students.

- *Cinema* quiz. The estimation of the reference rank μ does not coincide with the real rank even if the chronological order is correct for three films of four. However, the accuracy level of students knowledge in cinema is equal to $\pi = 0.723$. The students seem to have better knowledge in football than in cinema.

- *Rugby*. The ISR model estimation on the Rugby data set suggests a ranking between these four nations: during this time Scotland were the best, then Wales, England and finally Ireland. But the low value of the probability π (0.681) means that the confidence in this ranking is not very high.

- *Words*. The high value of π (0.879) for the Fligner and Verducci's Word data set shows that the questioned students overall share the same opinion for the association with the word Idea: Thought is the most associated followed by Theory, Dream, Attention and finally Play which is the least associated.

- *Sports*. The reference rank $(1,3,2,4,5,7,6)$ estimated for the ISR model reflects a preference of the students at the University of Illinois for collective sports: Baseball, Basketball and Football are at the top three places while individual sports are at the end of the ranking: Cycling, Jogging and Swimming. Tennis, which is intermediate between a collective sport and an individual sport, is rationally ranked between these two groups.

- *Election*. This last data set calls for an interesting remark: except the first two candidates, it suggests that all other candidates tends to be ordered similar to their initial order in the list. Note that we considered only 4% of the voters in our study, those who "bothered" to rank every one of the 14 candidates. Probably most of them have the common trait that they did not care who is ranked after the first two candidates since, in the Irish voting system, two-thirds of voters see their first choice elected. They may express such an indifference beyond the second rank by choosing the initial order in the list of candidates whereas the remaining 96% of voters prefer not to complete the rest of the list. Notice also that about 45% of these remaining 96% of voters select only one or two candidates. A more comprehensive study including partial ranks would be obviously required to analyse more precisely the results of these elections.

From the Mallows Φ parameters point of view, most results are highly consistent with ISR: main modal ranks are identical and the dispersion parameter λ is also well correlated with π, though λ is more abstract and could be less easy to understand by a practitioner. Only the modal rank of the last two data sets (Sports and Election) differs: for the Sports data set, Mallows Φ model classifies Tennis inside the collective sports collection instead of being put at the borderline of collective and individual sports, as ISR does; For the Election data set, Mallows Φ model provides the same first two candidates as ISR but in the reverse order and, in addition, all other candidates are very different between both methods.

5 Discussion

In this paper we propose that rank data could be considered as the result of a paired comparisons sort algorithm, where the possibility of wrong comparisons exists and occurs randomly according to a Bernoulli model. It opens a new way to propose many distributions on rankings, all of them benefiting from very meaningful parameters (the reference rank μ and the probability π of good paired comparison). In order to minimize the number of paired comparisons, and consequently number of potential wrong comparisons, the insertion sort algorithm has been kept in this paper for its optimality when $m \leq 10$ since it is expected to be a frequent limit in "human rankings". The resulting distribution, the so-called ISR, has been established and many desirable properties were being pointed out. In addition, the latent variable interpretation of model allows the derivation of specific EM and SEM-Gibbs algorithms which can be easily accelerated by drastically reducing the number of potential reference ranks μ to consider. Although our approach is not able to deal with very large data sets like web pages rankings or some kinds of biological data, these computational gains allow to deal with usual data sets of moderate size typically provided by a "human ranking" process.

In fact, the ISR model can be considered as the precursor to a wider family of ranking models. The insertion sort algorithm was kept in this work from some optimality arguments. However, it is possible to easily set up a new model by changing it. For instance, we can use a selection sort algorithm instead, or any other standard algorithms [**?**, see]]Knu1973. Obviously, properties of the new model obtained in this way would need to be established again. In each case, it could also be possible to weaken the assumption that π remains constant all along the sorting process, for instance to model the tiredness of the judge during the sorting process.

Another interesting prospect initialized by the present work is the possibility to include some information about the initial ranking y in the model and its corresponding estimation. Indeed, in questionnaires this initial order is often known and it is useful information which can be naturally used by our class of models. It is also possible to consider some more diffuse information about y, for instance to ignore the exact y value but to know that all y are the same for all questionnaires (realistic situation for many ranks coming from quiz studies), or other realistic variants [2]. In the same spirit, the model is also flexible enough to take into account the usual behaviour that some individuals may rank the items from the beginning to the end and others to do the opposite.

Although the ISR is unimodal (as many other distributions for ranks), multimodality can be easily taken into account through mixture of ISR distributions after leading a specific identifiability study. For instance, we can think that in our football quiz, girls and boys responses will probably not follow the same distribution, as it is suggested by very low estimated p-value [17]. This extension is natural since several mixtures of rank data models have already been considered with success to treat heterogeneity of rank population: mixture of Mallows Φ models [28], mixture of Plackett-Luce or Benter models [13, 15, 16] and more recently mixtures of weighted distance-based models [24].

Finally, there is also a need to adapt our models to other situations than full rank data. This approach needs to be extended to other types of ranks, frequently encountered in practice, as partially or incomplete ranked data [22, 20, 23], which would be very useful

to further analyse the Irish election data set, tied data or even ranks resulting from multiple preference responses.

Acknowledgment

We thank Pr. Brendan Murphy and Dr. Claire Gormley for providing the Irish election data set. We also thank the reviewers for their contributions to greatly improve this work.

A. Additional notations useful for the proofs

Additional notations are needed in the proofs displayed in the appendices below. All these notations are illustrated in detail through an example in Table 5. In addition, to clarify them, the calculation for step $j = 2$ (for instance) is detailed at the end of the present appendix.

In the following, $j = 1,\ldots,m$ denotes the step in the sorting algorithm consisting in ranking the object $\mathscr{O}_{y_j}$.

$A(x,y) = \sum_{j=1}^{m} A_j(x,y)$ and $A_j(x,y) = A_j^-(x,y) + A_j^+(x,y)$ designate the total number of *all* paired comparisons respectively for the whole process and for the step j. In this definition, $A_j^-(x,y)$ is the number of *all* comparisons of the current object $\mathscr{O}_{y_j}$ with the objects already ranked (according to x) before it (if they exist) and $A_j^+(x,y)$ indicates if the current object $\mathscr{O}_{y_j}$ is compared, at the j step of the sorting, with the object ranked in x just after it. Formally, $A_j^-(x,y)$ corresponds to the cardinal of $\mathscr{A}_j^-(x,y) = \{i : x_{y_i}^{-1} < x_{y_j}^{-1}, 1 \leq i < j\}$ which is the set of the indices of the presentation order y for which the already sorted objects $\mathscr{O}_{y_1},\ldots,\mathscr{O}_{y_{j-1}}$ are ranked in x before the current object $\mathscr{O}_{y_j}$. In a similar way, $A_j^+(x,y)$ is the cardinal of the set $\mathscr{A}_j^+(x,y) = \{i : i = \arg\min_{1\leq i'<j}\{i' : x_{y_{i'}}^{-1} > x_{y_j}^{-1}\}\}$ which corresponds to the index of the rank y designating the object sorted in x just after $\mathscr{O}_{y_j}$ among the already sorted objects $\mathscr{O}_{y_1},\ldots,\mathscr{O}_{y_{j-1}}$, if it exists. This set has at most one element.

$G(x,y,\mu) = \sum_{j=1}^{m} G_j(x,y,\mu)$ and $G_j(x,y,\mu) = G_j^-(x,y,\mu) + G_j^+(x,y,\mu)$ are the total number of *good* paired comparisons respectively for the whole process and for the step j. Formally, $G_j^-(x,y,\mu) = \sum_{i\in\mathscr{A}_j^-(x,y)} \delta_{y_i y_j}(\mu)$ is the number of *good* comparisons (according to μ) of the current object $\mathscr{O}_{y_j}$ with the objects already ranked before it (if they exist), where $\delta_{ii'}(\mu) = \mathbf{1}\{\mu_i^{-1} < \mu_{i'}^{-1}\}$ is equal to 1 if $\mathscr{O}_i$ is correctly ranked before $\mathscr{O}_{i'}$ (according to μ), 0 otherwise $(i,i' = 1,\ldots,m, i \neq i')$. In a similar way, $G_j^+(x,y,\mu) = \sum_{i\in\mathscr{A}_j^+(x,y)} \delta_{y_j y_i}(\mu)$ is the indicator of *good* comparison (according to μ) of the current object $\mathscr{O}_{y_j}$ with the object already ranked just after it (if it exists).

Detail of step $j = 2$ to be read in conjunction with Table 5

- $\mathscr{A}_2^- = \{\}$: the new object 3 (to be sorted) has been compared to the already sorted object 1 but 1 is not sorted *before* 3 in the final ordering x;

Table 5: An example to illustrate the notations with $\mu = (1,2,3)$, $y = (1,3,2)$ and $x = (3,1,2)$. The notation $x^{(j)}$, defined in B, means the ranking of the j first objects in y in the order imposed by x

step j	unsorted	sorted	$\mathscr{A}_j^-$	$\mathscr{A}_j^+$	A_j^-	A_j^+	A_j	G_j^-	G_j^+	G_j
start	$y = \boxed{1}\boxed{3}\boxed{2}$	-	-	-	-	-	-	-	-	-
1	$\boxed{3}\boxed{2}$	$x^{(1)} = \boxed{1}$	$\{\}$	$\{\}$	0	0	0	0	0	0
2	$\boxed{2}$	$\boxed{3} \overset{?}{\leftrightarrow} \boxed{1}$ $x^{(2)} = \boxed{3}\boxed{1}$	$\{\}$	$\{1\}$	0	1	1	0	0	0
3	-	$\boxed{2} \overset{?}{\leftrightarrow} \boxed{3}\boxed{1}$ $\boxed{3}\boxed{2} \overset{?}{\leftrightarrow} \boxed{1}$ $x = \boxed{3}\boxed{1}\boxed{2}$	$\{3,1\}$	$\{\}$	2	0	2	1	0	1
							$A = 3$			$G = 1$

- $\mathscr{A}_2^+ = \{1\}$: the new object 3 (to be sorted) has been compared to the already sorted object 1 and 1 is sorted *after* 3 in the final ordering x;
- $A_2^- = 0$ and $A_2^+ = 1$: number of comparisons listed respectively in $\mathscr{A}_2^-$ and $\mathscr{A}_2^+$;
- $A_2 = 1$: total number of comparisons to sort the new object 3.
- $G_2^- = 0$ and $G_2^+ = 1$: number of *good* comparisons listed respectively in $\mathscr{A}_2^-$ and $\mathscr{A}_2^+$;
- $G_2 = 1$: total number of *good* comparisons to sort the new object 3.

B. Building the ISR distribution

The goal of this appendix is to prove that (2.1) corresponds to the stochastic insertion sort algorithm with probability π of good paired comparison, and independence between the paired comparisons.

Proof. Let $x^{(j)}$ be the ordering of the first j ($1 \leq j \leq m$) objects in y in the order imposed by x (so $x^{(m)} = x$). An example of this notation is in Table 5. Thus, there exists following relationship between $x^{(j)}$ and $x^{(j-1)}$:

$$x^{(j)} = (x_1^{(j-1)}, \ldots, x_{A_j^-(x,y)}^{(j-1)}, y_j, x_{A_j^-(x,y)+1}^{(j-1)}, \ldots, x_{j-1}^{(j-1)}).$$

Equation (2.1) is now proved by induction on j. It is true for $j = 1$ while there is only one object y_1 to sort: $\mathrm{p}(x^{(1)}|y;\mu,\pi) = 1$. Since the result of the ranking $x^{(j)}$ from $x^{(j-1)}$ is the result of $A_j(x,y)$ independent Bernoulli experiments of parameter π, then, conditionally to $x^{(j-1)}$, the probability of $x^{(j)}$ is

$$\mathrm{p}(x^{(j)}|x^{(j-1)},y;\mu,\pi) = \pi^{G_j(x,y,\mu)}(1-\pi)^{A_j(x,y)-G_j(x,y,\mu)}.$$

We conclude the proof by noticing that

$$\mathrm{p}(x^{(j)}|y;\mu,\pi) = \mathrm{p}(x^{(j)}|x^{(j-1)},y;\mu,\pi)\mathrm{p}(x^{(j-1)}|y;\mu,\pi),$$

from the following implied relationship between events: $x^{(j)} \Rightarrow x^{(j-1)}$. □

C. Mathematical statement of the ISR properties and related proofs

In the following, composition $\tau \circ x$ will be noted shortly τx for any τ and x in $\mathscr{P}_m$.

Proposition 3. *(Uniformity for $\pi = \frac{1}{2}$.) For all $x, \mu \in \mathscr{P}_m$, $p(x;\mu,\frac{1}{2}) = m!^{-1}$.*

Proof. Let e be the identity permutation of $\mathscr{P}_m$. Firstly using Lemma 3 of D and then using the fact that $\mathrm{p}(.|e;\mu,\frac{1}{2})$ is a probability distribution on $\mathscr{P}_m$, we have

$$\mathrm{p}(x;\mu,\tfrac{1}{2}) \propto \sum_{y\in\mathscr{P}_m} \mathrm{p}(x|y;\mu,\tfrac{1}{2}) = \sum_{y\in\mathscr{P}_m} \mathrm{p}(y^{-1}x|y^{-1}y;\mu,\tfrac{1}{2}) = \sum_{y\in\mathscr{P}_m} \mathrm{p}(y^{-1}x|e;\mu,\tfrac{1}{2}) = 1.$$

□

Proposition 4. *(Mode μ.) For all $x \neq \mu \in \mathscr{P}_m$ and $\pi > \frac{1}{2}$, $p(\mu;\mu,\pi) > p(x;\mu,\pi)$.*

Proof. Using the fact that $\{\pi > \frac{1}{2} \Leftrightarrow \pi > 1-\pi\}$, $x \neq \mu$ and then Lemma 2, we obtain:

$$m!\,\mathrm{p}(x;\mu,\pi) < \sum_{y\in\mathscr{P}_m} \pi^{A(x,y)} = \sum_{y\in\mathscr{P}_m} \pi^{A((\mu x^{-1})x,(\mu x^{-1})y)} = \sum_{y'\in\mathscr{P}_m} \pi^{A(\mu,y')} = m!\,\mathrm{p}(\mu;\mu,\pi).$$

The last equality comes from the fact that $A(\mu,y') = G(\mu,y',\mu)$. □

Corollary 1. *(Anti-mode $\bar{\mu}$.)For all $x \neq \bar{\mu} \in \mathscr{P}_m$ and $\pi > \frac{1}{2}$, $p(\bar{\mu};\mu,\pi) < p(x;\mu,\pi)$.*

The proof is symmetrical to that of Proposition 4.

Proposition 5. *(Link between μ and π.) For all $x, \mu \in \mathscr{P}_m$, $p(\mu;\mu,\pi) - p(x;\mu,\pi)$ is an increasing function of $\pi \geq \frac{1}{2}$.*

Proof. Noting $\Delta(\pi) = \mathrm{p}(\mu;\mu,\pi) - \mathrm{p}(x;\mu,\pi)$, $\partial\Delta(\pi)/\partial\pi$, we can written

$$\frac{\partial\Delta(\pi)}{\partial\pi} = \frac{1}{m!}\sum_{y\in\mathscr{P}_m} \left\{A(\mu,y)\pi^{A(\mu,y)-1} - G(x,y,\mu)\pi^{G(x,y,\mu)-1}(1-\pi)^{A(x,y)-G(x,y,\mu)}\right\} + c,$$

where c is a non-negative term independent from π. Since $\pi \geq \frac{1}{2}$, we deduce that

$$G(x,y,\mu)\pi^{G(x,y,\mu)-1}(1-\pi)^{A(x,y)-G(x,y,\mu)} \leq G(x,y,\mu)\pi^{A(x,y)-1}.$$

Using the fact that $A(\mu,y) \geq G(x,y,\mu)$, we deduce that $\partial\Delta(\pi)/\partial\pi \geq 0$. □

Proposition 6. *(Symmetry.) For all $x, \mu \in \mathscr{P}_m$ and all $\pi \in [0,1]$, $p(x;\bar{\mu},1-\pi) = p(x;\mu,\pi)$.*

Proof. Using Lemma 4, we can write:

$$\mathrm{p}(x;\bar{\mu},1-\pi) \propto \sum_{y\in\mathscr{P}_m} \pi^{A(x,y)-(A(x,y)-G(x,y,\mu))}(1-\pi)^{A(x,y)-G(x,y,\mu)} \propto \mathrm{p}(x;\mu,\pi).$$

□

Proposition 7. *(Identifiability.) The* ISR *distribution is identifiable since* $\pi > \frac{1}{2}$.

Proof. The identifiability problem can concern parameters π and/or μ.

- First, there exists no couple $(\mu,\mu') \in \mathscr{P}_m^2$ with $\mu \neq \mu'$ such that $\mathrm{p}(x;\mu,\pi) = \mathrm{p}(x;\mu',\pi)$ for any $x \in \mathscr{P}_m$ and any $\pi > \frac{1}{2}$. Indeed, choosing $x = \mu$, from Lemma 5 we have $\mathrm{p}(\mu;\mu,\pi) \neq \mathrm{p}(\mu;\mu',\pi)$.
- Second, for a given $\mu \in \mathscr{P}_m$, assume there exists $\pi \neq \pi'$ such that $\mathrm{p}(x;\mu,\pi) = \mathrm{p}(x;\mu,\pi')$ for any $x \in \mathscr{P}_m$. In particular, for $x = \mu$, in the proof of Lemma 5 we obtained that $G(x,y,x) = A(x,y)$, thus $\sum_{y\in\mathscr{P}_m} \pi^{A(\mu,y)} = \sum_{y\in\mathscr{P}_m} \pi'^{A(\mu,y)}$. The strictly increasing function $p \mapsto \pi^n$ on the interval $[\frac{1}{2},1]$ for all $n \in \mathbb{N}^*$ ensures that $\pi = \pi'$.
- Assume finally there exists $(\mu,\mu') \in \mathscr{P}_m^2$ with $\mu \neq \mu'$ and $\pi < \pi'$ such that $\mathrm{p}(x;\mu,\pi) = \mathrm{p}(x;\mu',\pi')$ for any $x \in \mathscr{P}_m$. In the proof of Lemma 5, it is also obtained that $G(x,y,\mu) < A(x,y)$ when $x \neq \mu$, thus

  $$\mathrm{p}(x|y;\mu,\pi) < \pi^{A(x,y)} < \pi'^{A(x,y)} = \mathrm{p}(x|y;x,\pi'),$$

  and then by averaging over all y in $\mathscr{P}_m$ gives $\mathrm{p}(x;\mu,\pi) < \mathrm{p}(x;x,\pi')$. Choosing $x = \mu'$ ensures the identifiability of the ISR model.

□

D. Lemmas

Lemma 1. *Let* $\tilde{e} = (2,1,3,\ldots,m)$ *be the permutation inverting the first two elements. For all* $x,y,\mu \in \mathscr{P}_m$ *and* $\pi \in [0,1]$, $p(x|y;\mu,\pi) = p(x|y\tilde{e};\mu,\pi)$.

Proof. We use notations $x^{(j)}$ that have already been introduced in B. The key point of the proof is to notice that the first two objects in y lead to the same paired comparison at the second step of the sorting process whatever is their order in y, so $\mathrm{p}(x^{(2)}|y\tilde{e},\pi) = \mathrm{p}(x^{(2)}|y,\pi)$. Combining this result with the fact that $\mathrm{p}(x|x^{(2)},y\tilde{e},\pi) = \mathrm{p}(x|x^{(2)},y,\pi)$, since $\tilde{e}$ only affects the first two objects, this concludes the proof. □

Lemma 2. *For all* $x,y,\tau \in \mathscr{P}_m$, $A(x,y) = A(\tau x,\tau y)$.

Proof. First we prove that $A_j^-(x,y) = A_j^-(\tau x,\tau y)$. For any $j = 1,\ldots,m$, we have (notice that i is always such that $1 \leq i < j$)

$$\begin{aligned} A_j^-(\tau x,\tau y) &= \#\{i : (\tau x)^{-1}_{(\tau y)_i} < (\tau x)^{-1}_{(\tau y)_j}\} = \#\{i : (x^{-1}\tau^{-1}\tau y)_i < (x^{-1}\tau^{-1}\tau y)_j\} \\ &= \#\{i : (x^{-1}y)_i < (x^{-1}y)_j\} = \#\{i : x^{-1}_{y_i} < x^{-1}_{y_j}\} = A_j^-(x,y). \end{aligned}$$

Using the fact that $A_j^+(x,y) = \mathbf{1}\{A_j^-(x,y) + 1 \leq j-1\}$ we also deduce that $A_j^+(x,y) = A_j^+(\tau x,\tau y)$. Consequently, $A_j(x,y) = A_j(\tau x,\tau y)$ and, so, $A(x,y) = A(\tau x,\tau y)$. □

Lemma 3. *For all* $x,y,\mu,\tau \in \mathscr{P}_m$, $p(x|y;\mu,\frac{1}{2}) = p(\tau x|\tau y;\mu,\frac{1}{2})$.

Proof. When $\pi = \frac{1}{2}$, we obtain by using Lemma 2

$$\mathrm{p}(\tau x|\tau y;\mu,\tfrac{1}{2}) = \left(\tfrac{1}{2}\right)^{A(\tau x,\tau y)} = \left(\tfrac{1}{2}\right)^{A(x,y)} = \mathrm{p}(x|y;\mu,\tfrac{1}{2}).$$

□

Lemma 4. *For all* $x,y,\mu \in \mathscr{P}_m$ $G(x,y,\overline{\mu}) = A(x,y) - G(x,y,\mu)$.

Proof. Let $\overline{e}$ be the permutation of total inversion previously introduced in Section 2.3 and $i,i' = 1,\ldots,m, i \neq i'$. We first prove that $G_j^-(x,y,\overline{\mu}) = A_j^-(x,y) - G_j^-(x,y,\mu)$. Using successively the fact that $\overline{\mu} = \mu\overline{e}$, $\overline{e} = \overline{e}^{-1}$, $\{i < i' \Leftrightarrow \overline{e}_i > \overline{e}_{i'}\}$ and $i \neq i'$, we have

$$\begin{aligned}\delta_{ii'}(\overline{\mu}) &= \mathbf{1}\{(\mu\overline{e})_i^{-1} < (\mu\overline{e})_{i'}^{-1}\} = \mathbf{1}\{\overline{e}^{-1}_{\mu_i^{-1}} < \overline{e}^{-1}_{\mu_{i'}^{-1}}\} = \mathbf{1}\{\overline{e}_{\mu_i^{-1}} < \overline{e}_{\mu_{i'}^{-1}}\} \\ &= \mathbf{1}\{\mu_i^{-1} > \mu_{i'}^{-1}\} = 1 - \mathbf{1}\{\mu_i^{-1} < \mu_{i'}^{-1}\} = 1 - \delta_{ii'}(\mu),\end{aligned}$$

and then

$$G_j^-(x,y,\overline{\mu}) = \sum_{i \in \mathscr{A}_j^-(x,y)} (1 - \delta_{y_i y_j}(\mu)) = A_j^-(x,y) - G^-(x,y,\mu).$$

In a similar manner, we can prove that $G_j^+(x,y,\overline{\mu}) = A_j^+(x,y) - G_j^+(x,y,\mu)$. The proof follows immediately from these two results. □

Lemma 5. *For all* $x,\mu \in \mathscr{P}_m$, $x \neq \mu$ *and* $\pi > \frac{1}{2}$, $p(x;\mu,\pi) < p(x;x,\pi)$.

Proof. First note that $G(x,y,\mu) < A(x,y)$ for $\mu \neq x$. Since $\{\pi > \frac{1}{2} \Leftrightarrow 1-\pi < \pi\}$, we deduce for $\mu \neq x$ that $\mathrm{p}(x|y;\mu,\pi) < \pi^{A(x,y)}$. Also note that $G(x,y,x) = A(x,y)$, thus $\mathrm{p}(x|y;x,\pi) = \pi^{A(x,y)}$. Consequently, we have $\mathrm{p}(x|y;\mu,\pi) < \mathrm{p}(x|y;x,\pi)$ and the proof is concluded by averaging over all possible presentation orders y in $\mathscr{P}_m$. □

Lemma 6. *For all* $\mu,y \in \mathscr{P}_m$, $m-1 \leq A(\mu,y) \leq m(m-1)/2$.

Proof. Left bound: there is no comparison when the first element arises and at least one comparison for each of the $m-1$ other elements. Right bound: there is still no comparison when the first element arises and at most $j-1$ comparisons when for the jth new object to rank, so $A(\mu,y) \leq \sum_{j=1}^{m}(j-1) = m(m-1)/2$. □

E. Quiz data sets

References

[1] W. Benter. Computer-based horse race handicapping and wagering systems: A report. In W.T. Ziemba, V.S. Lo, and D.B. Haush, editors, *Efficiency of racetrack betting markets*. London: Academic Press, 1994.

Table 6: Quiz answers of the 40 students

Cinema		Football	
ordering	frequency	ordering	frequency
$(4,3,2,1)$	10	$(1,2,4,3)$	20
$(4,2,3,1)$	9	$(1,4,2,3)$	12
$(3,2,4,1)$	4	$(2,4,1,3)$	2
$(3,4,2,1)$	3	$(3,1,4,2)$	2
$(1,3,2,4)$	2	$(3,4,2,1)$	2
$(1,3,4,2)$	2	$(3,2,1,4)$	1
$(2,3,1,4)$	2	$(4,2,1,3)$	1
$(3,1,4,2)$	2	other	0
$(1,2,3,4)$	1		
$(2,3,4,1)$	1		
$(2,4,3,1)$	1		
$(3,2,1,4)$	1		
$(4,1,2,3)$	1		
$(4,3,1,2)$	1		
other	0		

[2] C. Biernacki and J. Jacques. Modèles génératifs de rangs relatifs à un algorithme de tri par insertion. In *42th Journées de Statistique organisée par la Société Française de Statistique*, Marseille, France, 2010.

[3] U. Böckenholt. Applications of Thurstonian models to ranking data. In *Probability models and statistical analyses for ranking data (Amherst, MA, 1990)*, volume 80 of *Lecture Notes in Statist.*, pages 157–172. Springer, New York, 1993.

[4] R.A. Bradley and M.E. Terry. Rank analysis of incomplete block designs. I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[5] G. Celeux and J. Diebolt. The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, 2(1):73–82, 1985.

[6] D. E. Critchlow. *Metric methods for analyzing partially ranked data*, volume 34 of *Lecture Notes in Statistics*. Springer-Verlag, Berlin, 1985.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[8] B. Efron and R.J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.

[9] P.D. Feigin and M. Alvo. Intergroup diversity and concordance for ranking data: an approach via metrics for permutations. *Ann. Statist.*, 14(2):691–707, 1986.

[10] M.A. Fligner and J.S. Verducci. Distance based ranking models. *J. Roy. Statist. Soc. Ser. B*, 48(3):359–369, 1986.

[11] M.A. Fligner and J.S. Verducci. Multistage ranking models. *J. Amer. Statist. Assoc.*, 83(403):892–901, 1988.

[12] A. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Matching Intelligence*, 6:721–741, 1984.

[13] I.C. Gormley and T.B. Murphy. Analysis of Irish third-level college applications data. *J. Roy. Statist. Soc. Ser. A*, 169(2):361–379, 2006.

[14] I.C. Gormley and T.B. Murphy. A latent space model for rank data. In *Proceedings of the 23th International Conference on Machine Learning*, Pittsburgh, PA, 2006.

[15] I.C. Gormley and T.B. Murphy. Exploring voting blocs within the irish electorate: A mixture modeling approach. *J. Amer. Statist. Assoc.*, 103(483):1014–1027, 2008.

[16] I.C. Gormley and T.B. Murphy. A mixture of experts model for rank data with applications in election studies. *Annals of Applied Statistics*, 2(4):1452–1477, 2008.

[17] J. Jacques and C. Biernacki. "model-based clustering for rank data based on an insertion sorting algorithm". In *17th Rencontres de la Société Francophone de Classification*, La Réunion, 2010.

[18] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[19] M.G. Kendall and B.B. Smith. On the method of paired comparisons. *Biometrika*, 31:324–345, 1940.

[20] P. Kidwell, G. Lebanon, and W.S. Cleveland. Visualizing incomplete and partially ranked data. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1356–1363, 2008.

[21] D.E. Knuth. *Sorting and Searching: Volume3. The art of Computer Programming*. Addison-Wesley, Massachusetts, 1973.

[22] G. Lebanon and J. Lafferty. Cranking: Combining rankings using conditional models on permutations. In *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia, 2002.

[23] G. Lebanon and Y. Mao. Non-parametric modeling of partially ranked data. *J. Mach. Learn. Res.*, 9:2401–2429, 2008.

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

[25] R.D. Luce. *Individual choice behavior: A theoretical analysis*. John Wiley & Sons Inc., New York, 1959.

[26] C.L. Mallows. Non-null ranking models. I. *Biometrika*, 44:114–130, 1957.

[27] J.I. Marden. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London, 1995.

[28] T.B. Murphy and D. Martin. Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, 41(3-4):645–655, 2003.

[29] R.L. Plackett. The analysis of permutations. *J. Roy. Statist. Soc. Ser. C Appl. Statist.*, 24(2):193–202, 1975.

[30] G.L. Thompson. Generalized permutation polytopes and exploratory graphical methods for ranked data. *Ann. Statist.*, 21(3):1401–1430, 1993.

[31] G.L. Thompson. Probability models and statistical analyses for ranking data. chapter Graphical techniques for ranked data, pages 294–298. Springer-Verlag, New-York, 1993.

[32] L.L. Thurstone. A law of comparative judgment. *Psychological Review*, 79:281–299, 1927.