International Journal of Business Analytics and Intelligence
6 (2) 2018, 02-05
http://publishingindia.com/ijbai/

Submitted: 10 June, 2018
Accepted: 20 July, 2018

Perspective: When Data are Ranks - Analysis of Rank Data

Arnab Kumar Laha*

In many real life situations respondents are asked to rank order a set of items based on their preferences. This can happen in selection interviews where a set of candidates have to be rank ordered (say, from best to worst) regarding their suitability for a job or position or in boardroom discussions where different alternative investment proposals have to be ranked based on their risk-reward profiles. In many market research studies respondents are asked to rank order a set of items with respect to their possibility of buying them. Thus rank data occur quite commonly in our daily life.

As each expert has his/her own criteria (often unexpressed or latent) for ranking a set of items, the rankings given by different experts may differ making it difficult to arrive at a "generally agreed" rank order. In selection interview panels, it's not uncommon to find experts who differ substantially with one another regarding the rank order of the interviewed candidates. To resolve these differences, ad-hoc methods are often used such as giving points according to the ranks. As an example think of a selection interview that has five candidates who appear for an interview. After the interview each expert rank order the five candidates based on their judgment. After this an ad-hoc scoring process is adopted in which the experts' rankings are converted to scores. A possible way can be that if an expert marks a candidate to be the best s/he is given five points, the next best is given four points and so on. At the end, the total score received by a candidate is computed as a sum of the scores obtained from each expert and the final rank is arrived at based on these total scores.

Can we avoid ad-hoc procedures such as the one described above and rely on scientific methods to arrive at the "generally agreed" (a.k.a. consensus) rank order? In other words, are there methods for analyzing rank data that arrive at the "generally agreed" rank order following a scientifically valid procedure? Over the years many methods have been proposed for analysing rank data. In this article we briefly discuss a few of them.

Let the number of items to be ranked by experts be k and let there be n experts. For simplicity of discussions let us assume k=3 and n=5 and think of a situation where three candidates appear for an interview having a panel consisting of five expert members. Now, at the end of the interviews each expert provides a ranking of the three candidates in terms of their suitability with 1 being the best and 3 being the worst. Thus, each expert gives an ordering of the three candidates in terms of their suitability. If an expert ranks the candidate B as "best", A as "next best" and C as the "worst" the ranking is denoted as $\begin{pmatrix} A & B & C \\ 2 & 1 & 3 \end{pmatrix}$. Thus each ranking can be thought of as a permutation of the elements of the set {1,2,3}. In particular, $\begin{pmatrix} A & B & C \\ 2 & 1 & 3 \end{pmatrix}$ can be written as the permutation (2 1 3). Suppose the rankings given by the five experts are (2 3 1), (2 1 3), (3 2 1), (3 1 2) and (1 3 2). What would be the "generally agreed" ranking in such a situation?

A possible solution to the above problem is provided by the "Kemeny ranking". The method is based on a notion of distance between permutations. Let and be two permutations. A possible distance is the Kemeny distance (a.k.a. Kemeny and Snell distance) which is defined as

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

$$d_k(\pi_1,\pi_2) = \frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}|x_{\pi_1}(i,j) - x_{\pi_2}(i,j)| \text{ where}$$

$$x_{\pi_s}(i,j) = \begin{cases} 1 & \text{if item } i \text{ is preferred over item } j \text{ in} \pi_s \\ -1 & \text{if item } j \text{ is preferred over item } i \text{ in} \pi_s \\ 0 & \text{if item } i \text{ and } j \text{ are tied in} \pi_s \end{cases}$$

(Kemeny and Snell, 1962). Let π_1 = (2 1 3) and π_2 = (1 1 2) . For ease of notation let the items A, B and C be denoted as 1, 2 and 3 respectively. Then we have

(i,j)	(1,2)	(1,3)	(2,3)
$x_{\pi_1}(i,j)$	-1	1	1
$x_{\pi_2}(i,j)$	1	-1	-1

Note that $x_{\pi_1}(i,j) = -x_{\pi_2}(i,j)$ and $x_{\pi_s}(i,j) = 0$. Thus $d_k(\pi_1,\pi_2) = 6$. In general, it can be easily seen that in case of rankings without any ties $d_k(0, 0)$ is always even.

Another distance function defined on the set of permutations is the Kendall-□ distance d_{ken}. For any two permutations π_1 and π_2, $d_{ken}(\pi_1, \pi_2)$ is defined as the number of "discordances" between the permutations π_1 and π_2. Formally, it is the number of elements in the set $D=\{(i,j):(\pi_1(i)< \pi_1(j) \text{ and } \pi_2(i) > \pi_2(j) \text{ or } (\pi_1(i) > \pi_1(j) \text{ and } \pi_2(i) < \pi_2(j))\}$ where $1 \le i < j \le k$. If π_1= (2 1 3) and π_2= (3 1 2) then D = {(1,2),(1,3),(2,3)} and $d_{Ken}(\pi_1, \pi_2) = 3$.

Let S_k be the set of all permutations of the k items. For each permutation $\pi \in S_k$ let $D_k(\pi) = \sum_{i=1}^{n} d_k(\pi_i,\pi)$ where be the rankings given by the n experts. The Kemeny ranking (a.k.a. Kemeny median) is defined as that permutation for which $D_k(\pi) = \min_{\pi \in S_k} D_k(\pi)$. Let us now compute the Kemeny median of the five expert rankings given above. Denote the five expert rankings as . Then we get (with), π_1 = (2,3,1), π_2 = (2,1,3), π_3 = (3,2,1), π_4 = (3,1,2) and π_5 = (1,3,2). Then we get $\left(\text{with } d_k \frac{1}{2} d_k\right)$

π	$d_k(\pi_1,\pi)$	$d_k(\pi_2,\pi)$	$d_k(\pi_3,\pi)$	$d_k(\pi_4,\pi)$	$d_k(\pi_5,\pi)$	$d_k(\pi)$
(1, 2, 3)	2	1	3	2	1	22
(1, 3, 2)	3	2	2	1	0	16
(2,1,3)	1	0	2	3	2	16
(2,3,1)	0	1	1	2	3	14
(3,1,2)	2	3	1	0	1	14
(3,2,1)	1	2	0	1	2	12

Thus the Kemeny median ranking for this example is (3,2,1) = (*C, B, A*) .

Let us consider a real-life example. Every year Financial Times (FT) comes out with its rankings of business schools from all over the world. We consider the FT rankings of the business schools for the years 2016, 2017 and 2018 derive from them the relative rankings of six Asian business schools which are given in the Table 1 below. Here 1 indicates the best, 2 the next best and so on. We are interested in finding the "generally agreed" or "overall" rank for these six schools based on this data.

	Table - I		
	Rank 2018	*Rank 2017*	*Rank 2016*
National University of Singapore Business School	1	1	4
Indian School of Business	2	2	3
Indian Institute of Management Ahmedabad	3	3	1
Shanghai Jiao Tong University: Antai	4	4	5
Indian Institute of Management Bangalore	5	6	6
CUHK Business School	6	5	2

Since here k=6, contains 6! = 720 elements. Hence it is not possible to do the computations by hand as for the earlier example. An R program can be easily developed for computing the Kemeny median of these rankings. The Kemeny ranking comes out to be (1, 2, 3, 4, 6, 5). FT also reports the average rank obtained by an institution over a three year period. In Table 2 we report the relative rankings derived from the FT average ranks and the Kemeny ranking.

	Ranking Based on Average FT Rank	*Kemeny Ranking*
National University of Singapore Business School	1	1
Indian School of Business	2.5	2
Indian Institute of Management Ahmedabad	2.5	3
Shanghai Jiao Tong University: Antai	5	4
Indian Institute of Management Bangalore	6	6
CUHK Business School	4	5

Sometimes in surveys, where there are many items to be ranked, instead of the full rankings the respondents are asked to provide the top-m (or bottom-m) items. For example from a list of business books the respondents may be asked to provide the top-3 books based on their usefulness for a MBA student. These rankings are referred to as partial rankings. The task now is to derive the rankings of all items in the list from the available partial rankings.

The method followed for obtaining the "generally agreed" ranking from partial rankings data uses the idea of tied ranking. For illustration, let there be 10 items on a list and suppose that the respondents are asked only to provide their top-3 items. If a respondent states (a, b, c) as her top-3 items, then all the other seven items are considered as tied at rank 4. With this modification it is now possible to compute the Kendall-distance d_{ken} (0, 0). Then the "generally agreed" ranking is obtained by minimising the criterion function $D(\pi) = \sum_{i=1}^{n} d_{ken}(\pi_i, \pi)$ where π_i, $1 \leq i \leq n$, are the observed partial rankings $\pi \in S_{10}$. Since S_K contains $k!$ elements and $k! \_ \sqrt{2\pi e}^{-k} k^{k+0.5}$ grows rapidly with increase in computation of "generally agreed" rank becomes computationally very expensive. This is particularly a more important issue when dealing with partial rankings as the total number of items to be ranked is typically large in this case.

Over the years several probability models for ranking data has been discussed in the literature. Here we discuss briefly a probability model for complete ranking data based on a distance measure (Mallows, 1957). Such models often take the form $P(\pi) = C(\lambda)e^{-\lambda d(\pi,\pi_0)}, \pi \in S_k$ where $\pi_0 \in S_k$ and $\lambda \geq 0$ are parameters. π_0 is called the modal ranking λ and is called the dispersion parameter. $C(\lambda)$ is the normalising constant that ensures $\sum_{\pi \in S_k} P(\pi) = 1$. If λ is large then the distribution of ranks is tightly clustered around π_0 whereas if λ is close to 0 then the distribution of ranks is close to uniform. Given a ranking dataset $\{\pi_1, ..., \pi_n\}$, the likelihood can be easily obtained as

$$L(\lambda, \pi_0) = C(\lambda)^n e^{-\lambda \sum_{i=1}^{n} d(\pi_i, \pi_0)}$$

As usual the MLEs (λ, π_0) of the parameters can be obtained by maximising $L(\lambda, \pi_0)$ over all possible values of (λ, π_0). In this context it may be noted that $\sum_{i=1}^{n} d(\pi_i, \pi_0) = \min_{\pi_0 \in S_k} \sum_{i=1}^{n} d(\pi_i, \pi_0)$. Thus if d_k is used as the distance measure then π_0 is the Kemeny ranking.

An alternative approach is to view the observed ranks as perturbations of the modal rank π_0 i.e. $\pi_i = \sigma_i \, 0 \, \pi_0$ where σ_i are i.i.d. S_k valued random variables and 0 denotes the composition of two permutations. A possible distribution on σ_i can be the Multinomial distribution $M(1; p_1, ..., p_{k!})$. Because the number of parameters in such models increases rapidly with increase in the number of items k, Bayesian analysis is often useful here. For more details the reader may see Laha and Dongaonkar (2012) and Laha et al. (2017).

References

Kemeny, J. G., & Snell, L. J. (1962). *Preference ranking: An axiomatic approach, in mathematical models in the social sciences*. Ginn., New York, 9-23.

Laha, A. K., & Dongaonkar, S. (2012). Bayesian analysis of rank data using SIR. In SenGupta, A. editor, *Advances in Multivariate Statistical Methods*, chapter 19, pp. 327-335. World Scientific.

Laha, A., Dutta, S., & Roy, V. (2017). A novel sandwich algorithm for empirical Bayes analysis of rank data. *Statistics and Its Interface*, *10*(4), 543-556.

Mallows, C. L. (1957). Non-null ranking models. *I. Biometrika*, *44*(1/2), 114-130.