COMPARISON OF PREDICTED AND REALIZED RANKINGS:

A SIMPLISTIC LIKELIHOOD APPROACH

by

David Hinkley

Technical Report No. 247

April, 1975

University of Minnesota
Minneapolis, Minnesota

Abstract

Some parametric models for stochastic permutations of integers are discussed in relation to comparison of predicted and realized rankings. Part of the purpose of the paper is to respond to a challenge by a rank correlation afficionado, I. D. Hill (1974), and to re-analyse Hill's data.

Key words: Randomness; Comparison of models; Rank correlation; Logistic model; Wilcoxon statistic; Prediction; Soccer.

1. Introduction.

In a recent article, Berry (1975) has discussed a novel parametric model for stochastic permutations of (1,2,...,N), parametrized so as to produce trends in the permutations, except in the null case when permutations are completely randomly generated. The model is intended to apply to a situation where rank correlation analysis is customarily applied, namely the comparison of two sets of rankings of N individuals or objects. One particular motivation for Berry's work was a challenge to the likelihood and Bayes "schools" by I. D. Hill (1974) in an article comparing predicted and realized ranks of British soccer teams.

My purpose here is to examine parametric models which I think might better represent the prediction situation, and which deserve attention as serious competitors. The models involve a simple notion of group prediction, complemented by a model for the order of permutations. These are described in Section 3, and fitted to the soccer data in Section 4. Section 2 contains a brief review of Berry's model with some further discussion of its properties.

2. Stochastic Permutations: Berry's Model

The data under consideration constitute a single permutation of the integers $(1,2,\ldots,N)$, each possible permutation having equal probability $(N!)^{-1}$ only under a null hypothesis of completely random stochastic permutation. If an arbitrary permutation is $I = (I_1,\ldots,I_N)$, we then define the vector $X = (X_1,\ldots,X_N)$ to be the positions of $1,2,\ldots,N$ in I; that is,

$$X_j = k \text{ if and only if } I_k = j .$$

A stochastic model for such permutations that tends to produce trends (correlation with the natural order $1,\ldots,N$), proposed and discussed by Berry (1975), is

$$P(X_1 = x_1,\ldots,X_n = x_N) = q(x_1;\theta)\prod_{j=2}^{N} q(x_j|x_1,\ldots,x_{j-1};\theta) , \qquad (2.1)$$

where

$$q(x_1;\theta) = \frac{x_1^{-\theta}}{\sum_a a^{-\theta}} , \quad q(x_j|x_1,\ldots,x_{j-1};\theta) = \frac{x_j^{-\theta}}{\sum_{\substack{a\neq x_s \\ s=1,\ldots,j-1}} a^{-\theta}} , \ (j = 2,\ldots,N)$$

and $\sum_a$ denotes summation over $1,2,\ldots,N$. Key properties of the model (2.1) are as follows.

(i) $\theta = 0$ corresponds to completely random permutation;

(ii) $\theta > 0$ corresponds to a tendency for the x_j to be increasing, the tendency being stronger for large θ;

(iii) $\theta < 0$ corresponds to a tendency for the x_j to be decreasing, the tendency being stronger for large $|\theta|$;

(iv) The most probable sequences are $x = (1,2,\ldots,N)$ when $\theta > 0$ and $x = (N,N-1,\ldots,1)$ when $\theta < 0$.

The joint probability (2.1) will be thought of as a likelihood, and is best expressed as

$$\text{lik}_B(\theta|x) = \frac{(N!)^{-\theta}}{\prod_{j=0}^{N} \left(\sum_{a \neq x_1,\ldots,x_j} a^{-\theta} \right)} , \quad x_0 = 0 . \tag{2.2}$$

This likelihood is unimodal, so that the maximum likelihood estimate is well-defined, but (2.2) is asymmetric, in the sense that $\text{lik}_B(\theta|x = (c_1,\ldots,c_N)) \neq \text{lik}_B(-\theta|x = (c_N,\ldots,c_1))$. Qualitatively, when $\theta > 0$, the trend in x is strongest at the beginning of the sequence and weakest at the end of the sequence.

The latter remarks are particularly relevant since the model (2.2) applies to data for which the classical statistical analysis is based on rank correlations. Both Spearman's rho and Kendall's tau coefficients are symmetric statistics. Recall that

$$\text{Spearman rho} = 1 - \frac{6 \sum (x_j - 1)^2}{N(N^2 - 1)}$$

and that

$$\text{Kendall tau} = \frac{S}{\frac{1}{2}N(N-1)} ,$$

where
$$S = 2 \sum_{j=1}^{N-1} \sum_{k=j+1}^{N} h(x_k - x_j) - \tfrac{1}{2}N(N-1) \tag{2.3}$$

with $h(u) = 1$ and 0 according as $u \geq 0$ and $u < 0$. As an example, consider the two sequences

$$x = (5,4,3,2,1,6,7,8,9,10) \quad \text{and} \quad x' = (1,2,3,4,5,10,9,8,7,6) .$$

The values of rho and tau are respectively 0.76 and 0.56 for both sequences, but the likelihood functions $\text{lik}_B(\theta|x)$ and $\text{lik}_B(\theta|x')$

have maxima at 1.0 and 3.4 respectively, and relative likelihoods (relative to maximum values) at $\theta = 0$ are 0.15 and 0.04 respectively. Of course it may be that in practice one would want the extra emphasis on "early" integers that $\mathrm{lik}_B(\theta|x)$ possesses, but the lack of correspondence with rank correlation is unfortunate for comparison of "likelihood inference" and "sampling theory inference". A symmetric parametric model for X is defined in Section 3.

Neither θ nor rank correlations have strong physical interpretation without appeal to an underlying continuous-variate from which the ranks are defined. Since any systematic permutation is, by definition, non-random the search for a physically interpretable parametrization is probably futile.

3. Alternative Models for Prediction Applications

3.1. Motivation.

The specific focus of our discussion is the following problem, a special case of which occupied Hill (1974): A set of N individuals are to take a test, following which they will be ranked. Prior to the test an expert predicts the post-test ranks, and we wish to compare the two sets of ranks so as to determine whether the results of the test are predictable, and if so to what degree.

The particular data we have for analysis are predicted and actual end-of-season ranks for each of six British soccer leagues in the 1971-2 season. A typical example is

$$x = (6,2,7,3,5,9,15,8,1,4,19,16,11,18,17,10,14,21,12,13,22,20) \qquad (3.1)$$

which is the set of actual positions of English Football League Division 1 teams predicted to finish in first place, second place, etc.

Now it is not unnatural to suppose that the prediction is carried out in at least two stages, the first of which is a classification into categories such as "good", "medium" and "poor", or simply "above average" and "below average". Further stages in the prediction process would then lead to the ranks attached to individuals. For simplicity, and to maintain a degree of objectivity, we shall assume that there is an initial classification into two groups of roughly equal size, following which individuals are ranked within groups.

After defining suitable models for relationships between predicted and actual ranks, our interest will naturally be in determining whether the predictors can do more than classify into groups. The remainder of this section is devoted to defining plausible models and some comment on their analysis.

3.2. Models for Group Identification.

We shall suppose, without loss of generality, that predictors place individuals $1,\dots,m$ in group G^+("above average") and individuals $m+1,\dots,N$ in group G^-("below average"), and that individual j has predicted rank j. The final order of the individuals defines the positions of the N individuals as $X_1,\dots,X_N$, but the success of group classification is determined solely by the unsequenced values of $X_1,\dots,X_m$.

It is convenient to introduce the notation

$$Y_j = \begin{cases} 1 & j \in (X_1,\dots,X_m) \\ 0 & j \in (X_{m+1},\dots,X_N) \end{cases} \qquad (j = 1,\dots,N) ,$$

with the restriction $\sum_1^N Y_j = m$. There are $\binom{N}{m}$ possible sequences $(y_1,\dots,y_N)$, each equally likely if group classification is random. The simplest general model for Y is the logistic probability function

$$\text{lik}_L(\lambda|x) = P(Y_1 = y_1,\dots,Y_N = y_N) = \frac{\exp(\lambda \sum_1^m y_j)}{\sum_{s=0}^{\min(m,n)} \binom{m}{s}\binom{n}{m-s}\exp(\lambda s)}, \qquad (3.2)$$

where $n = N - m$. The emphasis here is on how many of those individuals predicted to be in G^+ do not actually belong to G^+. The model corresponds to a 2 X 2 contingency table, the independence hypothesis corresponding to $\lambda = 0$.

A more practical model for prediction is obtained by paying attention to which individuals are incorrectly classified (if any). We take as our second model

$$\mathrm{lik}_W(w|x) = p(Y_1 = y_1,\dots,Y_N = y_N) = \frac{\exp(-\omega \sum y_j x_j)}{\binom{N}{m} \sum_t \exp(-\omega t) p_{m,n}(t)} \qquad (3.3)$$

where $p_{m,n}(\cdot)$ is the probability distribution of the Wilcoxon rank-sum statistic $W = \sum_1^m X_j$ under the null hypothesis of complete randomness, which corresponds to $\omega = 0$. The exact evaluation of $p_{m,n}(\cdot)$ is straightforward using the recurrence relation

$$p_{m,n}(t) = \frac{m}{N}\, p_{m-1,n}(t-N) + \frac{n}{N}\, p_{m,n-1}(t)$$

together with the identities: $W = 0$ if $m = 0$, $W = \frac{m(m+1)}{2}$ if $n = 0$.

Since both $\sum Y_j$ and $\sum X_j Y_j$ are asymptotically normal for large m,n under complete randomness, large-sample approximations may be deduced from (3.2) and (3.3). These are

$$\mathrm{lik}_L(\lambda|x) \doteq \binom{N}{m}^{-1} \exp\left(\lambda \sum_1^m y_j - \lambda\frac{m^2}{N} - \tfrac{1}{2}\lambda^2 \frac{m^2 n^2}{N^2(N-1)}\right)$$

and

$$\mathrm{lik}_W(\omega|x) \doteq \binom{N}{m}^{-1} \exp\left(-\omega \sum y_j x_j + \omega\mu_T - \tfrac{1}{2}\omega^2 \sigma_T^2\right)$$

where $\mu_T = \frac{m(N+1)}{2}$ and $\sigma_T^2 = \frac{mn(N+1)}{12}$.

Unfortunately it turns out that these approximations are not accurate for the sizes of samples ($N \le 24$) considered in Section 4.

3.3. Models for the Sequence X.

We suppose that after prediction of G^+ and G^- the order within each group is predicted. To examine whether or not this prediction of order is successful, aside from group classification, we require probability models for the actual sequences $(X_1,\dots,X_m)$ and $(X_{m+1},\dots,X_N)$

conditional on which individuals finally belong to G^+ and G^-. That is, we must model the final term in the equation

$$P(X = x) = P(Y = y)P(X = x|Y = y)$$

to arrive at a model for the complete sequence $X = (X_1,\ldots,X_N)$.

Let $r^+ = (r_1^+,\ldots,r_m^+)$ be the ranks of $x_1,\ldots,x_m$ within G^+, and let $r^- = (r_1^-,\ldots,r_n^-)$ be the ranks of $x_{m+1},\ldots,x_N$ within G^-. For example if x is given by (3.1), then

$$r^+ = (6,2,7,3,5,9,10,8,1,4,11) .$$

We assume, with loss of generality, that R^+ and R^- are independently distributed. One possible model for X conditional on Y is then

$$P(X = x|Y = y) = \text{lik}_B(\theta^+|r^+)\text{lik}_B(\theta^-|r^-) , \qquad (3.4)$$

where $\text{lik}_B(\theta|\cdot)$ is Berry's model defined in (2.2). We may be particularly interested in comparing predictability of orders within G^+ and G^-, in which case comparison of θ^+ and θ^- would be of interest. See, however, the discussion of Berry's model in Section 2; possibly r^- should be replaced by $n - r^- + 1$ to emphasise the predictability of the worst individuals, rather than those near the middle. I have not studied this possibility in the context of the soccer data.

The lack of symmetry of Berry's model (Section 2), in particular its incoherence with rank correlation statistics, leads me to consider an alternative symmetric model for stochastic permutations. Here the device used in defining (3.3) is used again, namely embedding a classical test statistic, such as Spearman's rho, as sufficient statistic in an exponential family. The difficulty with this is calculation of the denominator.

Since the exact distribution of Spearman's rho is very difficult to compute, we shall use Kendall's tau, so that corresponding to (2.2) we define

$$\mathrm{lik}_C(\gamma|x) = \frac{\exp(-\gamma S)}{\Sigma p_N^*(t)\exp(-\gamma t)} \tag{3.5}$$

where S is given by (2.3) and $p_N^*(\cdot)$ is the probability distribution of S under complete randomness of permutation. It is well-known (Kendall, 1962, p. 67) that

$$p_{k+1}^*(t) = \frac{1}{k+1}\sum_{j=0}^{k} p_k^*(t - k + 2j)$$

with $t = -\frac{1}{2}k(k-1), -\frac{1}{2}k(k-1)+2, \ldots, \frac{1}{2}k(k-1)$ as the support of $p_k^*(\cdot)$ and $p_1^*(0) = 1$.

For the two-group model we must apply (3.5) twice and then our alternative to (3.4) is

$$P(X = x|Y = y) = \mathrm{lik}_C(\gamma^+|r^+)\mathrm{lik}_C(\gamma^-|r^-) \; . \tag{3.6}$$

In summary, we now have two possible models for group classification, viz. (3.2) and (3.3), and two possible models for the order of X within each group, viz. (3.4) and (3.6). The combinations into overall models for X will be denoted with mnemonic suffices as

$$\mathrm{lik}_{LC}(\lambda,\gamma^+,\gamma^-|x) = \mathrm{lik}_L(\lambda|x)\mathrm{lik}_C(\gamma^+|r^+)\mathrm{lik}_C(\gamma^-|r^-) \; ,$$

etc. It should be noted that such three-parameter models do not overlap the one-parameter models $\mathrm{lik}_B(\theta|x)$ and $\mathrm{lik}_C(\gamma|x)$ except for the null case of complete randomness.

3.4. Inferential Use of the Likelihoods.

The composite models with three parameters, such as $\text{lik}_{WB}(\omega,\theta^{+},\theta^{-}|x)$, are competitors to $\text{lik}_{B}(\theta|x)$. Within such a three-parameter family there is a natural hierarchy of hypotheses, for example $H_0:\theta^{+} = \theta^{-} = \omega = 0$, $H_1:\theta^{+} = \theta^{-} = 0$, $\omega \neq 0$, $H_3:\theta^{+},\theta^{-},\omega$ all non-zero. Particular attention will focus on H_1, and the likelihood under H_1 vis à vis the corresponding one-parameter likelihood, $\text{lik}_{B}(\theta|x)$, since overall preference for H_1 suggests that prediction does little more than identify groups.

In comparisons of two models each with one free parameter, the ratio of maximized likelihoods is a natural measure of which model fits better. This is not entirely satisfactory, but a loose justification is as follows: The two single-parameter functions $\text{lik}_{WC}(\omega,0,0|x)$ and $\text{lik}_{C}(\gamma|x)$ may be thought of as belonging to one likelihood family, continuous in the sense that both components go through the common uniform null distribution. Therefore that component with larger maximum indicates the likelihood estimate of best model in the combined family. Unfortunately the combined family is not smooth (the two component families meet at a sharp angle) and the likelihood estimate may well be biased.

Comparison of a three-parameter likelihood such as $\text{lik}_{WC}(\omega,\gamma^{+},\gamma^{-}|x)$ with a one-parameter likelihood such as $\text{lik}_{C}(\gamma|x)$ is difficult, because the families have only one common point (H_0). In general one might complement the likelihoods with relevant prior distributions of parameters, but I find the task of assigning reasonable priors formidable here. For statistical problems such as comparison of alternative linear models, there is some justification for offsetting the ratio of maximized likelihoods by the factor $\exp\{\frac{1}{2}(d_{num} - d_{den})\}$, where d_{num} and d_{den} are dimensions of numerator and denominator parameter sets; see Kanemasu (1973). Personally I feel that a somewhat larger discount factor is required for N near 20, as in the soccer data.

I am not aware of a satisfactory account of pure likelihood model comparision; if a sampling-theory approach were taken, Cox's (1961) work would be relevant, but would require more effort to implement than our expert predictors probably merit!

One possible general approach was suggested to me by some comments of G. A. Barnard, namely that the observed likelihood ratio can be calibrated by determining the corresponding likelihood ratio based on neutral data. Two difficulties with this idea are, first, that the choice of neutral data is to some extent subjective and, second, that one individual may choose two sets of ostensibly neutral data which lead to different likelihood ratios. Nevertheless, I think the approach may be useful.

This discussion will prepare the reader for some rather ad hoc likelihood data analysis in the next section.

4. Model Fits for the Soccer Prediction Data.

A prime motivation for the discussion in Section 3 was the data analysed by Hill (1974). Briefly, for each of the top six British professional soccer leagues an expert panel of journalists predicted end-of-season team places within the league prior to the 1971-2 season. The actual places of teams with predicted ranks 1,2,...,n define the x vector discussed above for each of the six leagues. These vectors are given in Table 1.

Each of the models described in Sections 2 and 3 was fitted to each x vector, with group sizes $m = n$ in each but the last case, where $m = 10$, $n = 9$. The basic summary for each likelihood fit is the maximum likelihood relative to the likelihood at the null parameter value, which we denote simply by LR_A for models $A = B,C,L,W$. For models B and C fitted to G^+ and G^- as described in Section 3.3, the superscripts + and - are added. Thus, for example,

$$LR_C^+ = \frac{\sup \text{lik}(\gamma|r^+)}{\text{lik}(0|r^+)} \quad .$$

These likelihood ratios are given in Table 2.

Consider first the various two-group models. With one clear exception (SLD1), the Wilcoxon model better describes the success of group classification. Thus, generally, as expected, when misclassification occurs it is not random, but rather in favor of the middle-rank teams. Success at group classification is high, with the exceptions of FLD4 and, possibly, FLD2.

TABLE 1

End-of-season league positions (x) of soccer teams within each of Football League Divisions 1 through 4 (FLD1 - FLD4) and Scottish League Divisions 1 and 2 (SLD1 - 2), season 1971-1972. Teams in order of pre-season predictions.

League	Size N	x
FLD1	22	6 2 7 3 5 9 15 8 1 4 19 16 11 18 17 10 14 21 12 13 22 20
FLD2	22	2 6 12 7 19 9 14 3 10 11 13 5 1 4 20 18 16 15 21 8 17 22
FLD3	24	1 7 10 4 6 3 17 2 11 23 14 5 21 13 8 18 22 12 16 9 20 15 19 24
FLD4	24	11 3 20 17 8 16 2 12 9 21 24 1 5 15 13 4 23 6 7 19 14 10 18 22
SLD1	18	1 3 2 8 15 4 9 6 5 12 18 7 10 14 13 11 17 16
SLD2	19	4 1 9 10 2 11 7 5 16 18 6 14 3 12 19 17 13 8 15

TABLE 2

Ratios of maximum likelihood to null parameter likelihood for component models fitted to Table 1 data. [LR_A = sup $lik_A(\alpha|x)$ $lik_A(0|x)$ for models A = B,C,L,W with parameters $\alpha = \theta, \gamma, \lambda, \omega$; superscripts + and - indicate model fitted to G^+ and G^- respectively]

	FLD1	FLD2	FLD3	FLD4	SLD1	SLD2
LR_B	105	6.03	965	2.58	9097	15.3
LR_C	652	17.6	166	1.49	470	15.22
LR_L	95.3	2.20	20.0	1.38	303	3.43
LR_W	340	2.53	61.4	1.05	144	9.46
LR_B^+	1.09	3.40	2.11	1.41	8.50	3.90
LR_B^-	2.11	17.2	2.61	5.81	2.23	1.23
LR_C^+	1.68	3.11	2.18	1.01	3.05	3.05
LR_C^-	1.68	5.45	1.41	3.25	1.42	1.004
LR(W:B)	3.2	0.42	0.064	0.42	0.016	0.62
LR(W:C)	0.52	0.14	0.37	0.73	0.31	0.62

Also, by inspection of the values of LR_C^+ and LR_C^-, there is not much success at predicting the order within groups, particularly in FLD1, FLD3 and SLD2. Note that LR_B^+ and LR_C^+ values are usually comparable, with the exception of SLD1 (where the <u>real</u> "good" group is acknowledged to be teams 1,2 and 3).

Now compare the simple Wilcoxon group classification model with the <u>overall</u> trend models, for which we use ratios

$$LR(W:B) = \frac{lik_{WB}(\hat{\omega},0,0)}{lik_B(\hat{\theta})} = \frac{LR_W}{m!n!LR_B}$$

and

$$LR(W:C) = \frac{LR_W}{m!n!LR_C}$$

given in Table 2. Here the evidence is in favor of the Kendall tau model C as against W, but not strongly. The Berry model, which emphasises the top-ranked teams, is definitely inferior to the Wilcoxon model for FLD1. These results encourage me to believe that the major part of prediction success comes from group classification success, but there is a consistent minor ability to predict the order. Comparisons of full group classification models with overall trend models, using $LR(W:C)LR_C^+ LR_C^-$ and so on, do not indicate definite superiority of the former.

Following the remarks in Section 3.4 about calibrating likelihood ratios, it would probably not be possible to agree on a neutral x for distinguishing between W and C, but one possible candidate is

$$x = (1,N,2,N-1,\ldots,[\tfrac{1}{2}(N-1)] + 1) .$$

For the six values of N in our data, the values of LR(W:C) are 1.02, 1.02, 0.95, 0.95, 1.02, 0.96 and the corresponding values of

LR(W:B) are 0.10, 0.10, 0.07, 0.07, 0.16, 0.13. These suggest that the Wilcoxon model is as credible as the data values of LR(W:C) naively indicate, and more credible than the data values of LR(W:B) indicate.

Generally Berry's model does not fare well overall relative to the Kendall tau model, SLD1 being an exception for a reason already posited..

All models agree that FLD4 was unpredictable.

5. Summary.

Comparison of two sets of rankings is sometimes more meaningfully done in terms of groups of individuals rather than the individuals themselves. Some simple parametric models for permutations are available, notably those based on Kendall's tau statistic and Wilcoxon's group rank sum statistic. Of course similar results might be obtained by computing rank correlations within groups, and suitably comparing these with overall rank correlations. However, the use of likelihoods is probably more straightforward for interpretation purposes.

6. Acknowledgements.

I wish to thank Don Berry for patient listening and for not giving me a prior distribution, and Claudia Tysdal for continued excellent manuscript preparation. I am particularly grateful to S. K. Lee for his expert programming of all calculations. The support of a grant-in-aid from the Graduate School, University of Minnesota, was most helpful.

REFERENCES

Berry, D. (1975). On the presence of trends in "random" sequences. Univ. of Minnesota School of Statistics Tech. Report No. 239.

Cox, D. R. (1961). Tests of separate families of hypotheses. Proc. 4th Berk. Symp. Prob. and Stat. 1, 105-123.

Hill, I. D. (1974). Association football and statistical inferences. Appl. Statist. 23, 203--8.

Kanemasu, H. (1973). Posterior probabilities of candidate models in model discrimination. Tech. Report. No. 322, Dept. of Stat., Univ. of Wisconsin, Madison.

Kendall, M. G. (1962). Rank Correlation Methods. Griffin, London.