Computational Statistics and Data Analysis 54 (2010) 1672–1682

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Distance-based tree models for ranking data

Paul H. Lee *, Philip L.H. Yu

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 6 November 2008
Received in revised form 1 January 2010
Accepted 21 January 2010
Available online 1 February 2010

Keywords:
Decision tree
Ranking data
Distance-based model

ABSTRACT

Ranking data has applications in different fields of studies, like marketing, psychology and politics. Over the years, many models for ranking data have been developed. Among them, distance-based ranking models, which originate from the classical rank correlations, postulate that the probability of observing a ranking of items depends on the distance between the observed ranking and a modal ranking. The closer to the modal ranking, the higher the ranking probability is. However, such a model basically assumes a homogeneous population and does not incorporate the presence of covariates.

To overcome these limitations, we combine the strength of a tree model and the existing distance-based models to build a model that can handle more complexity and improve prediction accuracy. We will introduce a recursive partitioning algorithm for building a tree model with a distance-based ranking model fitted at each leaf. We will also consider new weighted distance measures which allow different weights for different ranks in formulating more flexible distance-based tree models. Finally, we will apply the proposed methodology to analyze a ranking dataset of Inglehart's items collected in the 1999 European Values Studies.

1. Introduction

Ranking data frequently occurs where judges (or individuals) are asked to rank a set of items, which may be types of soft drinks, political goals, candidates in an election, etc. By studying ranking data, we can understand judges' perception and preferences on the ranked alternatives.

Over the years, various statistical models for ranking data have been developed such as order statistics models, rankings induced by paired comparisons (for instance, Bradley–Terry Model), distance-based models and multistage models; see Critchlow et al. (1991) and Marden (1995) for more details of these models. Among these models, distance-based models have the advantages of being simple and elegant. However, they have received much less attention than what they should deserve, probably because they have two major weaknesses: the assumption of homogeneous population and difficulties in incorporating covariates. These have greatly limited the usefulness of the model.

Distance-based models (Fligner and Verducci, 1986) assume a modal ranking π_0 and the probability of observing a ranking π is inversely proportional to its distance from the modal ranking. The closer to the modal ranking π_0, the more frequent the ranking π is observed. Many distance measures have been proposed in the literature. Typical examples of distances are Kendall, Spearman and Cayley's distances; see Mallows (1957), Critchlow (1985) and Diaconis (1988). The models with Kendall's distance is sometimes referred as Mallows' ϕ-model (Mallows, 1957). The models consist of only two parameters, modal ranking π_0 and dispersion λ, and yet can provide a useful descriptive summary to a set of ranking data. Simplicity is obvious.

* Tel.: +852 96846294.
E-mail addresses: honglee@hku.hk (P.H. Lee), plhyu@hku.hk (P.L.H. Yu).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.01.027

The distance-based models assume a homogeneous population, all individuals will have a consensus view on the ranking of the items which is summarized by the modal ranking π_0. However, this may not be always true. Recently, Murphy and Martin (2003) extended the use of mixtures to distance-based models to describe the heterogeneity among the judges. By relaxing the homogeneous assumption, this leads to a significant improvement in the model, but the model still does not incorporate the presence of covariates.

There are quite a number of developments for including covariates. For example Beggs et al. (1981), Chapman and Staelin (1982), Hausman and Ruud (1987) and Train (2003) discussed the rank-ordered logit model (originated from order statistic model) which can incorporate covariates. Yu (2000) developed the multivariate order statistic models for ranking data, which can incorporate covariates as well. Gormley and Murphy (2008) adopted the mixture of experts model introduced by Jacobs et al. (1991) which allows covariates for mixture model, but again with an order statistic model. For distance-based models, the inability to incorporate covariates still remains a major inadequacy. This paper aim at developing a distance-based model by incorporating covariates, and hence addressing the heterogeneity population, by making use of a decision tree approach.

Decision trees are statistical models designed for classification and prediction problems. A decision tree is so called because the prediction rules generated from a set of the covariates can be displayed in a tree-like structure. Because of their ease of model interpretation, and the automatic detection of important covariates and interaction effects, tree-based models have been developed successfully in extending classical statistical models including logistic regression tree (Chan and Loh, 2004), Poisson regression tree (Chaudhuri et al., 1995), log-normal regression tree (Ahn, 1996) and generalized autoregressive conditional heteroscedastic (GARCH) tree (Audrino and Bühlmann, 2001). In this paper, we will combine a decision tree and a distance-based model so as to develop a more flexible distance-based model which can allow the presence of covariates.

The remainder of this paper is organized as follows. Section 2 reviews the distance-based models for ranking data and proposes the new weighted distance-based models. Section 3 proposes an algorithm of building distance-based tree models for ranking data. To illustrate the feasibility of the proposed algorithm, a simulation study and a case study of real data are presented in Sections 4 and 5 respectively. Finally, some concluding remarks are given in Section 6.

2. Distance-based models for ranking data

2.1. Distance-based models

Some notations are defined here for better description of ranking data. When ranking k items, labeled $1, \ldots, k$, a ranking $\boldsymbol{\pi}$ is a mapping function from $1, \ldots, k$ to $1, \ldots, k$, where $\pi(i)$ is the rank given to item i. For example, $\pi(2) = 3$ means that item 2 is ranked third.

Distance function is useful in measuring the discrepancy in two rankings. The usual properties of a distance function are:

- $d(\boldsymbol{\pi}, \boldsymbol{\pi}) = 0$,
- $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) > 0$ if $\boldsymbol{\pi} \neq \boldsymbol{\sigma}$,
- $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = d(\boldsymbol{\sigma}, \boldsymbol{\pi})$.

For ranking data, we require the distance, apart from the usual properties, to be right invariant, i.e. $d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = d(\boldsymbol{\pi} \circ \boldsymbol{\tau}, \boldsymbol{\sigma} \circ \boldsymbol{\tau})$, where $\boldsymbol{\pi} \circ \boldsymbol{\tau}(i) = \boldsymbol{\pi}(\boldsymbol{\tau}(i))$. This requirement makes sure relabeling of items has no effect on the distance.

Some popular distances are Spearman's rho, given by

$$R(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \left(\sum_{i=1}^{k} [\pi(i) - \sigma(i)]^2 \right)^{0.5}. \tag{1}$$

Spearman's rho square, given by

$$R^2(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} [\pi(i) - \sigma(i)]^2. \tag{2}$$

Spearman's footrule, given by

$$F(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} |\pi(i) - \sigma(i)|, \tag{3}$$

and Kendall's tau, given by

$$T(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i<j} I\{[\pi(i) - \pi(j)][\sigma(i) - \sigma(j)] < 0\}, \tag{4}$$

where $I\{\}$ is the indicator function. Apart from these distances, there are other distances for ranking data, and readers can refer to Critchlow et al. (1991) for details.

Diaconis (1988) developed a class of distance-based models,

$$P(\boldsymbol{\pi}|\lambda, \boldsymbol{\pi}_0) = \frac{e^{-\lambda d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}}{C(\lambda)}, \tag{5}$$

where $\lambda \geq 0$ is the dispersion parameter, and $d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ is an arbitrary right invariant distance. In particular, when we use Kendall's tau as the distance function, the model is called Mallows' ϕ-model (Mallows, 1957). Rankings nearer to the modal ranking $\boldsymbol{\pi}_0$ have a higher probability of occurrence and this is controlled by λ. The distribution of ranking will be more concentrated around $\boldsymbol{\pi}_0$ for smaller λ.

The close form for proportionality constant $C(\lambda)$ only exists for some distances. In principle, it can be solved numerically by summing the value $e^{-\lambda d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}$ over all possible $\boldsymbol{\pi}$. This numerical calculation is time-consuming, as the computational time actually increases exponentially with the number of items. For more details readers can refer to Fligner and Verducci (1986).

Suppose we have a ranking dataset $D = \{\boldsymbol{\pi}_i, i = 1, \ldots, n\}$. If the modal ranking $\boldsymbol{\pi}_0$ is known, the maximum likelihood estimator (MLE) $\hat{\lambda}$ of the distance-based model can be found by solving the equation

$$\frac{1}{n}\sum_{m=1}^{n} d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_0) = E_{\hat{\lambda}, \boldsymbol{\pi}_0}[d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)]. \tag{6}$$

The Left hand side (LHS) and right hand side (RHS) of the equation can be interpreted as the observed mean distance and expected distance respectively.

The MLE can be found numerically because LHS is a constant and RHS is a strictly increasing function of $\hat{\lambda}$. For the ease of solving, we re-parameterize λ with ϕ where $\phi = e^{-\lambda}$. The range of ϕ lies in $(0, 1]$ and $\hat{\phi}$ can be obtained using the method of bisection. Critchlow (1985) suggested applying the method with 15 iterations, which yields an error of 2^{-15}. Central Limit Theorem holds for the MLE $\hat{\lambda}_0$ which is shown by Marden (1995).

If the MLE $\hat{\boldsymbol{\pi}}_0$ is unknown, it can be computed as follows. The MLE $\hat{\boldsymbol{\pi}}_0$ minimizes the sum of distance, that is:

$$\sum_i d(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_0). \tag{7}$$

For large k, global search algorithm for MLE $\hat{\boldsymbol{\pi}}_0$ is not practical because the number of possible choices is too large. Instead, as suggested by Busse et al. (2007), a local search algorithm should be used.

2.2. ϕ-component models

Fligner and Verducci (1986) extended the distance-based models by decomposing the distance metric $d(\boldsymbol{\pi}, \boldsymbol{\sigma})$ into $k-1$ distance metrics,

$$d(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k-1} d_i(\boldsymbol{\pi}, \boldsymbol{\sigma}), \tag{8}$$

where the $d_i(\boldsymbol{\pi}, \boldsymbol{\sigma})$'s are independent. Both Kendall's tau and Cayley's distance can be decomposed in this form, and Fligner and Verducci (1986) developed two new classes of ranking model, called ϕ-component models and cyclic structure models, for the decomposition of Kendall's tau and Cayley's distance respectively. We will work with the ϕ-component models but not the cyclic structure models because Cayley's distance is not appropriate for our application.

Fligner and Verducci (1986) showed that Kendall's tau satisfies Eq. (8):

$$T(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{\pi_0(i)=1}^{k-1} V_{\pi_0(i)}, \tag{9}$$

where

$$V_{\pi_0(i)} = \sum_{\pi_0(j)=\pi_0(i)+1}^{k} I\{[\pi(i) - \pi(j)] > 0\}. \tag{10}$$

Here, V_1 represents the number of adjacent transpositions required to place the best item in $\boldsymbol{\pi}_0$ in the first position. V_2 is the number of adjacent transpositions required to place the second best item in $\boldsymbol{\pi}_0$ in the second position, and so on. Therefore, the ranking can be described as $k-1$ stages, V_1 to V_{k-1}, where $V_i = m$ can be interpreted as m mistakes made in stage i.

By applying dispersion parameter λ_i on stage V_i, Mallows' ϕ-model is extended to:

$$P(\boldsymbol{\pi}|\boldsymbol{\lambda}, \boldsymbol{\pi}_0) = \frac{e^{-\sum_{\pi_0(i)=1}^{k-1} \lambda_{\pi_0(i)} V_{\pi_0(i)}}}{C(\boldsymbol{\lambda})}, \tag{11}$$

where $\boldsymbol{\lambda} = \{\lambda_i, i = 1, \ldots, k-1\}$ and $C(\boldsymbol{\lambda})$ is the proportionality constant.

These models are named $k-1$ parameter models in Fligner and Verducci (1986), but were also named ϕ-component models in other papers (e.g. Critchlow et al., 1991). Mallows' ϕ-models are special cases of ϕ-component models when $\lambda_1 = \cdots = \lambda_{k-1}$.

The extension of distance-based models to $k-1$ parameters allows more flexibility in the model, but unfortunately, the symmetric property of distance is lost. Notice that the so-called "distance" in ϕ-component models can be expressed as

$$\sum_{\pi_0(i)<\pi_0(j)} \lambda_{\pi_0(i)} I\{[\pi(i) - \pi(j)] > 0\}, \tag{12}$$

which is obviously not symmetric, and hence it is not a proper distance measure. For example, in ϕ-component model, let $\boldsymbol{\pi} = (2, 3, 4, 1)$, $\boldsymbol{\pi}_0 = (4, 3, 1, 2)$. $d(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \lambda_1 V_1 + \lambda_2 V_2 + \lambda_3 V_3 = 3\lambda_1 + 0\lambda_2 + 1\lambda_3 \neq 1\lambda_1 + 2\lambda_2 + 1\lambda_3 = d(\boldsymbol{\pi}_0, \boldsymbol{\pi})$. The symmetric property of distance cannot be satisfied.

2.3. Weighted distance-based models

We propose an extension of distance-based model by replacing the (unweighted) distance with a new weighted distance measure, so that different weights can be assigned to different ranks.

Motivated from the weighted Kendall's tau correlation coefficient proposed by Shieh (1998), we define weighted Kendall's tau distance by

$$T_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i<j} w_{\pi_0(i)} w_{\pi_0(j)} I\{[\pi(i) - \pi(j)][\sigma(i) - \sigma(j)] < 0\}. \tag{13}$$

Note that this weighted distance satisfies all the usual distance properties, in particular, the symmetric property: $T_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\sigma}) = T_{\mathbf{w}}(\boldsymbol{\sigma}, \boldsymbol{\pi})$.

Other distance measures can be generalized to weighted distance in a similar manner as what we have done in generalizing Kendall's tau distance. Some examples are given below.

Weighted Spearman's rho square is

$$R^2_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} w_{\pi_0(i)} [\pi(i) - \sigma(i)]^2. \tag{14}$$

Weighted Spearman's rho is

$$R_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \left(\sum_{i=1}^{k} w_{\pi_0(i)} [\pi(i) - \sigma(i)]^2 \right)^{0.5}. \tag{15}$$

Weighted Spearman's footrule is

$$F_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\sigma}) = \sum_{i=1}^{k} w_{\pi_0(i)} |\pi(i) - \sigma(i)|. \tag{16}$$

Apart from the weighted Kendall's tau (Shieh, 1998) and weighted Spearman rho square (Shieh et al., 2000), there are many other weighted rank correlations proposed; see, for example, Tarsitano (2009).

Applying a weighted distance measure $d_{\mathbf{w}}$ to distance-based model, the probability of observing a ranking $\boldsymbol{\pi}$ under the weighted distance-based ranking model is

$$P(\boldsymbol{\pi}|\mathbf{w}, \boldsymbol{\pi}_0) = \frac{e^{-d_{\mathbf{w}}(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}}{C(\mathbf{w})}. \tag{17}$$

Generally speaking, if w_i is large, few people will disagree the item which ranked i in $\boldsymbol{\pi}_0$, because this disagreement will greatly increase the distance and hence probability of observing it will be very small. If w_i is close to zero, people have no preference about how the item which ranked i in $\boldsymbol{\pi}_0$ is ranked, because the change of its rank will not affect the distance at all.

3. Distance-based tree models for ranking data

Among many tree building strategies, the classification and regression tree (CART) procedure (Breiman et al., 1984) is the most popular one which consists of two stages: growing and pruning. The growing stage begins with the root node, which contains the entire learning sample. The root node is partitioned into two subgroups, referred to as child nodes. These same procedures can be applied to split the two child nodes, leading to the recursive partitioning process. (Breiman et al., 1984) The growing process continues until a stopping criterion is met. In the pruning stage, the cost-complexity pruning method is applied (Breiman et al., 1984) and the tree is chopped into a reasonable size for easy understanding and interpretation. This can avoid any overfit of data by the tree model.

Our proposed methodology for constructing distance-based tree models, following CART method, will be explained in this section. Pseudo-code for demonstrating how our algorithm works is provided in Fig. 1. We will also propose a goodness-of-fit measure for our tree models.

while #node > 1 **do** {Generating subtrees}
  $R(t)$ = deviance
  $g(t) = \frac{R(t)-R(T_t)}{|\tilde{T}_t|-1}$
  if min $g(t)$ **then**
    chopped nodes below this node
    add current tree to list
  end if
end while
return $T^0, T^1, ..., T^m$
for $i = 1$ to V **do** {Cross-validation}
  Data $\rightarrow L_v$ and L_v^C
  generate $T_v^0, T_v^1, ..., T_v^{m_v}$
  store $g(t_v^0), g(t_v^1), ..., g(t_v^m)$
  $i \leftarrow i + 1$
end for
for $i = 1$ to m **do**
  compute $R^{CV}(T^i)$
  store T^* which minimize $R^{CV}(T)$
  $i \leftarrow i + 1$
end for
for $i = 1$ to m **do** {1-SE rule}
  if $R^{CV}(T^i) \leq R^{CV}(T^*) + SE(R^{CV}(T^*))$ **then**
    select T^i as final model
  else
    $i \leftarrow i + 1$
  end if
end for

Fig. 1. Pseudo-code of tree pruning algorithm.

3.1. The growing stage

Suppose we have n observations $(\boldsymbol{\pi}_i, X_i)$, $i = 1, \ldots, n$, where X_i is a collection of m-dimensional covariates. Here, the X_i's can be categorical, interval or ordinal variables. In the following, we will illustrate how we can construct a distance-based tree model using these observations, with the aim of building a predictive model for our target $\boldsymbol{\pi}$ based on X.

The tree growing stage is a top–down algorithm. The root node, containing all observations, will be split into left child node N_L and right child node N_R by a splitting rule S, which comprises a splitting variable X^s and a splitting point. The child nodes will then be further split recursively according to splitting rules (other than S) until a certain stopping criterion is met. All nodes that cannot be further split are called terminal nodes, and the others are called internal nodes.

In selecting the splitting rule for each internal node, we shall choose the one that minimizes the weighted sum of the mean deviances ($D_L + D_R$) of two child nodes formed, i.e., $N_L D_L + N_R D_R$. For weighted distance-based model, the deviance of a particular node y with size N_y is

$$D_y = \frac{2}{N_y} \sum_{i=1}^{N_y} d_{\hat{\mathbf{w}}}(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_0) + \log C(\hat{\mathbf{w}}), \tag{18}$$

which equals $-\frac{2}{N_y} \times$ loglikelihood. Exhaustive search algorithm is used to determine the best splitting rule that gives the smallest weighted sum of mean deviances of the two child nodes. D in (18) will also be used in other tree models such as distance-based tree models.

Next, we turn to the issue of heterogeneity among judges. Not surprisingly, the difference of deviance between the mother node and the two child nodes is an indicator of the existence of heterogeneity. A larger decrease in deviance implies that the chosen splitting rule splits the data into two populations of larger difference. On the other hand, if the decrease in deviance is small, it means that the data in the two nodes follow two similar distributions.

However, the weighted mean deviance described above has one drawback. The weighted mean deviance criterion often results in a split which produces two unbalanced nodes, where one of them is small but pure. To avoid this problem, we will stop splitting when the size of a child node is smaller than $\frac{1}{10}$ of the size of its mother node. Besides, in order to avoid

overfitting, a node with sample size smaller than 200 (this figure is arbitrary, and is proven to work well by experience) will not be further split, and it will automatically become a terminal node.

3.2. The pruning stage

In the growing stage, we tend to build an overly large tree, hoping not to miss any important features of the tree. As a result, many abundant nodes will be created and interpretation will become difficult. Pruning is necessary to remove these abundant nodes.

The pruning procedure of our model follows the cost-complexity algorithm in CART (Breiman et al., 1984). It is named so as we want to find a tree model with both small error and size. The tree which minimizes

$$R(T) + g|\tilde{T}|, \tag{19}$$

will be chosen, where $R(T)$ and $g|\tilde{T}|$ are the penalization of large error and large tree size respectively. The pruning stage can be divided into two steps: (1) generating subtrees and (2) choosing the best subtree. We will discuss them in the following.

3.2.1. Generating subtrees

Before explaining how the subtrees are generated, here are some definitions:

- $R(T)$ is the cost function (or error) of a tree T. It is defined as the sum of deviance (Eq. (18)) for all of its terminal nodes. If T itself is a terminal node, its $R(t)$ will be the deviance of this node.
- T_t is the subtree with root t.
- $|\tilde{T}|$ is the number of terminal nodes in the tree T.
- $g(t)$ is the strength of the link from an internal node t and is defined as $\frac{R(t)-R(T_t)}{|\tilde{T}_t|-1}$. It can be viewed as the reduction of error if the node t is further split instead of stopped.

Note that when $g = g(t)$, it is indifferent for cutting the nodes under t or not.

After the growing stage, we arrive at a tree T^0. The pruning stage begins by calculating $g(t)$ for all internal nodes in T^0 and searching for the node t^1 with the minimum value $g(t^1)$. Next, we then cut all the descendant nodes under t^1 and turn t^1 into a terminal node. A pruned tree T^1 is then created. This process is repeated until T^0 is pruned to the root T^m. Afterwards, we obtain a sequence of nested trees $T^0 \supset T^1 \supset T^2 \supset \cdots \supset T^m$. Among the sequence, the tree T^y will be chosen as our final model if $g(t^y) \le g \le g(t^{y+1})$.

3.2.2. Choosing the best tree

After generating the trees $T^0, T^1, \ldots, T^m$, there are two ways to choose the best model. Breiman et al. (1984) suggested two methods. When the dataset is large, the independent testing dataset method can be used. Otherwise, the cross-validation method is suggested. In this paper, the latter method is applied.

To compute our selection criterion, the cross-validation deviance (DEV-CV), the n observations are divided randomly into V arbitrary equal-sized subsets $L_1, L_2, \ldots, L_V$. It is a common practice to take $V = 10$. For $v = 1, \ldots, V$, L_v is the vth validation dataset and its complement L_v^C is the vth training dataset. Based on the training dataset L_v^C, the vth sequence of nested trees $T_v^0 \supset T_v^1 \supset T_v^2 \supset \cdots \supset T_v^{m_v}$ is constructed. We then use the validation dataset L_v to compute the validation deviance. The validation deviance in any particular terminal node of tree T_v^y, D_v^y is given as $2\sum_{i=1}^{n_v} d_{\hat{\mathbf{w}}_v^y}(\boldsymbol{\pi}_i, \hat{\boldsymbol{\pi}}_{0v}^y) + \log C(\hat{\mathbf{w}}_v^y)$, where $\hat{\mathbf{w}}_v^y$ and $\hat{\boldsymbol{\pi}}_{0v}^y$ are estimated using training dataset L_v^C, n_v and $\boldsymbol{\pi}_i$ are the dataset size and the observed rankings of the validation dataset L_v respectively. Adding up the validation deviance of all terminal nodes, we get the validation deviance of tree T_v^y.

After computing the validation deviances for all nested trees $T_v^0, T_v^1, \ldots, T_v^{m_v}$, $v = 1, \ldots, V$, we can proceed to the next step, that is, evaluating the cross-validation deviance for each of the nested trees $T^0, T^1, \ldots, T^m$. Breiman et al. (1984) recommended using the geometric average $\sqrt{g(t^y)g(t^{y+1})}$ as the estimate of g^y, and hence the cross-validation deviance for tree T^y, $R^{CV}(T^y)$ is evaluated as

$$R^{CV}(T^y) = \frac{1}{V}\sum_{v=1}^{V} R\left(T_v\left(\sqrt{g(t_v^y)g(t_v^{y+1})}\right)\right) \tag{20}$$

where $T_v(g)$ is equal to the pruned subtree T_v^y such that $g(t_v^y) \le g \le g(t_v^{y+1})$.

Now we can return to the selection of the best subtree from $T^0, T^1, \ldots, T^m$. Let T^* be the tree with minimum cross-validation deviance. It sounds reasonable to select T^* as our final model. Since the position of the minimum $R^{CV}(T^*)$ is uncertain (Breiman et al., 1984), we adopt the 1-SE rule to select another tree T^{**} as our final model. T^{**} is the smallest subtree which satisfies

$$R^{CV}(T^{**}) \le R^{CV}(T^*) + SE(R^{CV}(T^*)). \tag{21}$$

Table 1
Summary of modal rankings.

Child node	π_0
$X_1 = 0$	1, 2, 3
$X_1 = 1$	3, 2, 1

Table 2
Simulation results (data generated from Mallows' ϕ-model).

Distribution of X_2	ϕ	X_1	X_2
Ber(0.5)	0.1	100	0
	0.3	100	0
	0.5	92	8
	0.7	76	24
	0.9	60	40
C(4)	0.1	100	0
	0.3	100	0
	0.5	74	28
	0.7	54	46
	0.9	17	83
O(21)	0.1	100	0
	0.3	100	0
	0.5	68	32
	0.7	39	61
	0.9	0	100

3.3. Model assessment

The cross-validation deviance $R^{CV}(T^{**})$ can measure the model prediction performance of our model. Besides, to assess the goodness-of-fit of the model, we use sum of squares Pearson residuals (χ^2) suggested by Marden (1995). χ^2 equals $\sum_i^{k!} r_i^2$, where $r_i = \frac{(O_i - E_i)}{\sqrt{E_i}}$ is the Pearson residual, O_i and E_i are observed and expected frequencies of ranking i.

However, if the size of some E_i are smaller than 5, the computed chi-square statistic will be biased. We are likely to encounter this problem when the size of the dataset is small and k is large. In this case, we suggest using the truncated sum of squares Pearson residuals criterion described by Erosheva et al. (2007).

4. Simulation study

A simulation study is carried out to compare the discriminatory power of the splitting criterion under different conditions. Suppose we have 100 observations in a ranked dataset of three items. The rankings are generated from Mallows' ϕ-model with parameters ϕ and π_0. X_1 is a Bernoulli random variable with $p = 0.5$, as shown in Table 1.

A noise variable X_2 is added to the dataset. Different values of ϕ and different distributions of X_2 are used to find out the discriminatory power of the split. The number of times the splitting variable X_1 or X_2 is selected out of 100 replications under various distribution are summarized in Table 2. Ber(p) denotes Bernoulli distribution with parameter p, C(n) denotes a uniform discrete distribution of n values and O(n) denotes a uniform ordinal distribution of n values.

The results show that, for every distribution of X_2, the probability of selecting the correct splitting variable increases when the parameter ϕ decreases. This is expected because a smaller ϕ implies that the data is more condensed around π_0, hence the two populations with $X_1 = 0$ and $X_1 = 1$ will become more distinct.

From Table 2, we also find that exhaustive search is biased towards selecting variables with more choices of splits. When X_1 and X_2 are both binary, the probability of correctly choosing X_1 is much higher than that of X_2. When X_2 has a C(4) distribution, there are a total of 7 possible splits, and X_2 is found to have a much higher probability of being selected, the highest $\left(\frac{83}{100}\right)$ occurring when $\phi = 0.9$. When X_2 has a O(21) distribution, the number of possible splits increases from 7 to 20, and the probability of selecting X_2 further increases. Even when $\phi = 0.7$, X_2 is selected frequently $\left(\frac{61}{100}\right)$.

Another similar simulation is conducted, where the data is generated using Spearman's rho distance-based model. The data is fitted using Mallows' ϕ-models as in the previous simulation. The results shown in Table 3 indicate that the selection of splitting variable is insensitive to the distance used in the model.

5. Application to the European Values Studies

In order to test the applicability of the distance-based tree models described in Section 3, we have made use of the ranked dataset obtained from the European Values Studies (EVS). It is a continuing, annual program of cross-national collaboration on surveys covering topics important for social science research. In this paper, we will examine the survey which was conducted in 1999 in 32 countries in Europe (Vermunt, 2004).

Table 3
Simulation results (data generated from Spearman's rho distance-based model).

Distribution of X_2	ϕ	X_1	X_2
Ber(0.5)	0.1	100	0
	0.3	100	0
	0.5	100	0
	0.7	95	5
	0.9	53	47
C(4)	0.1	100	0
	0.3	100	0
	0.5	99	1
	0.7	31	69
	0.9	0	100
O(21)	0.1	100	0
	0.3	100	0
	0.5	83	17
	0.7	2	98
	0.9	0	100

Table 4
Subject covariates for EVS data.

Covariate	Description/code	Type
Country	Group 1 = 1, Group 2 = 2, Group 3 = 3, Group 4 = 4	Nominal
Gender	Male = 1, Female = 2	Binary
Year of birth	Value ranges from 1909–1981	Interval
Marital status	Married = 1, Widowed = 2, Divorced = 3, Separated = 4, Never married = 5	Nominal
Employment status	Ordinal value ranges from 1–8	Ordinal
Household income	Ordinal value ranges from 1–10	Ordinal
Age of education completion	Interval value ranges from 7–50	Interval

Table 5
The four groups of countries.

Group 1	Group 2	Group 3	Group 4
Italy	Croatia	Luxembourg	Lithuania
Sweden	Belgium	Slovenia	Latvia
Denmark	Greece	Czechnia	Poland
Austria	France	Iceland	Estonia
Netherlands	Spain	Finland	Belarus
	Northern Ireland	West Germany	Slovakia
	Ireland	Portugal	Hungary
		Romania	Ukraine
		Malta	Russia
		East Germany	
		Bulgaria	

The total number of respondent in the survey is 1911. The respondents' covariates used in tree building are summarized in Table 4. Countries are categorized into four groups as suggested by Vermunt (2004); see Table 5 for the four groups of countries. We use 75% of the data for model building (1433 observations) and the remaining data (478 observations) for testing model performance.

The survey mainly focused on value orientations, attitudes, beliefs and knowledge concerning nature and environmental issues, and included the so-called Inglehart Index (Inglehart, 1977), a collection of four indicators of materialism/postmaterialism as well. Respondents were asked to pick the most important and the second most important political goals for their Government from the following four alternatives:

(A) Maintain order in nation.
(B) Give people more say in Government decisions.
(C) Fight rising prices.
(D) Protect freedom of speech.

Respondents can be classified into three value priority groups according to their top 2 choices. "Materialist" corresponds to individuals who give priority to (A) and (C) regardless of the ordering, whereas those who choose (B) and (D) will be classified as "postmaterialist". The rest will be classified as those who hold "mixed" value orientations.

Table 6
Loglikelihood, AIC and deviance for different models.

Model	Size	Log L	AIC	DEV-CV	DEV-testing
(Unweighted) Distance-based models					
Kendall's tau	1	−3387	6776	2.4913	2.8536
ϕ-component	1	−3296	6596	2.3015	2.5231
Rho square	1	−3365	6732	2.4920	2.8532
Rho	1	−3430	6862	2.4634	2.6810
Footrule	1	−3407	6816	2.4920	2.5281
Weighted distance-based models					
Weighted tau	1	−3252	6512	2.2722	2.4705
Weighted rho square	1	−3495	6994	2.8402	2.9930
Weighted rho	1	−3732	7468	2.6001	2.6362
Weighted footrule	1	−3762	7528	2.6267	2.6663
(Unweighted) Distance-based tree models					
Kendall's tau tree	5	−3460	6938	2.4578	2.3755
ϕ-component tree	3	−3271	6558	2.2854	2.3871
Rho square tree	4	−3451	6916	2.4599	2.6167
Rho tree	12	−3545	7136	2.4389	2.4521
Footrule tree	5	−3467	6952	2.4452	2.3906
Weighted distance-based tree models					
Weighted tau tree	5	−3074	6196	2.1766	**2.2562**
Weighted rho square tree	2	−3494	7006	2.3761	2.5377
Weighted rho tree	4	−3454	6946	2.4033	2.5332
Weighted footrule tree	4	−3495	7028	2.4222	2.5041

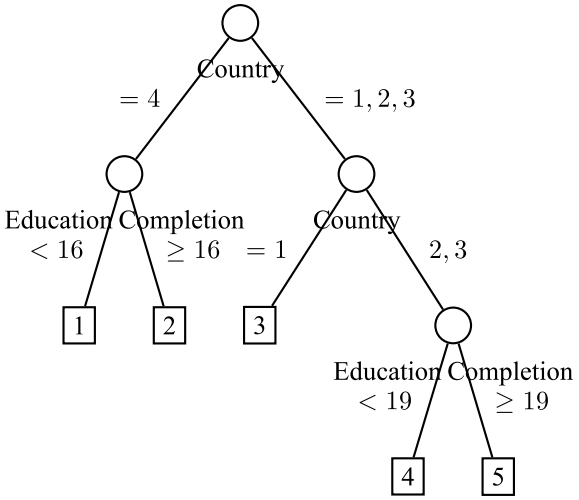

Fig. 2. Weighted tau tree model.

In this survey, respondents assigned ranks for only 2 out of the 4 available items. This type of rankings is called partial ranking and we need to modify the likelihood in order to fit the distance-based models to this dataset. Since we have no preference information about the non-ranked items in a partial ranking, it is natural to assume that all possible rankings that are compatible with the partial ranking are equally likely to be observed. Therefore, the probability of observing a partial ranking $\boldsymbol{\pi}^*$ is the sum of the probabilities of all possible complete ranking $\boldsymbol{\pi}$ that are compatible with $\boldsymbol{\pi}^*$.

For rankings with 4 items, the computation time was acceptable. We performed our model building in a computer with Pentium 4 CPU 3.40 GHz, and 1 GB of RAM. The tree building algorithm was completed in about 2 min and the pruning algorithm was completed in about 10 min.

Four classes of distance-based models are considered. Table 6 shows the comparison of different distance-based models and their corresponding tree models and their weighted distance versions. In terms of cross-validation deviance, tree-based extensions of distance-based models with different types of distances are all better than the original models. Weighted distance-based tree models perform better than ϕ-component tree models. The best among them is the weighted Kendall's tau tree model. Fig. 2 shows the weighted Kendall's tau tree. Table 7 shows the parameter estimates in the terminal nodes of weighted Kendall's tau tree model. The corresponding weighted Kendall's tau tree is shown for comparison. Table 8 shows

Table 7
Parameter estimates of the fitted weighted Kendall's tau tree model.

Model	Node	Size	Ordering of goals in π_0	w_1	w_2	w_3	w_4
Weighted Kendall's tau		1433	$A \succ B \succ C \succ D$	0.94	0.36	0.50	0.67
	1	114	$A \succ C \succ B \succ D$	0.69	1.24	0.10	2.24
	2	487	$A \succ B \succ C \succ D$	0.79	0.24	0.84	1.58
Weighted Kendall's tau tree	3	177	$A \succ C \succ D \succ B$	0.64	0.55	0.26	1.40
	4	383	$A \succ B \succ C \succ D$	1.33	0.20	0.52	0.65
	5	272	$A \succ B \succ D \succ C$	0.59	1.07	0.27	1.00

Table 8
Sum of squares Pearson residuals of root node and final nodes for different models, evaluated using testing data.

–	AB	AC	AD	BA	BC	BD	CA	CB	CD	DA	DB	DC	χ^2
Observed	76	97	41	57	38	34	51	23	11	23	22	5	
Expected (ϕ-component)	85.53	79.39	73.72	40.16	43.26	40.16	21.87	25.37	21.86	11.91	13.82	12.83	
Residual	−1.03	1.98	−3.81	2.66	−0.80	−0.97	6.23	−0.47	−2.32	3.21	2.20	−2.19	92.53
Expected (tau tree)	83.29	48.96	53.83	56.00	30.42	31.57	33.15	21.33	16.43	36.69	24.06	18.01	
Residual	−0.80	6.87	−1.75	0.13	1.38	0.43	3.10	0.36	−1.33	−2.26	−0.42	−3.07	79.16
Expected (weighted tau)	94.97	80.99	52.37	63.64	30.78	18.76	46.60	24.85	15.66	24.84	14.05	10.51	
Residual	−1.95	1.78	−1.57	−0.83	1.30	3.52	0.64	−0.37	−1.18	−0.37	2.12	−1.70	33.65
Expected (weighted tau tree)	88.99	90.52	44.44	65.03	32.54	19.86	44.71	23.72	14.09	27.49	17.66	8.29	
Residual	−1.37	0.68	−0.52	−1.00	0.96	3.17	0.94	−0.14	−0.82	−0.86	1.03	−1.14	19.29

the observed and expected frequencies of testing data and the χ^2 values for the four classes of models. The χ^2 values reveal that the weighted Kendall's tau tree model provides the best fit to the data.

In general, the two covariates, country and education level are important predictors of how Inglehart's items are ranked. Country is a more important predictor than age of education completion. For Group 4 countries (mainly former USSR countries), (A) and (C) are preferred and they can be classified as materialist type. For other country groups, (A) and (B) are preferred and they are the mixed type. For people who finish education before the age of 19, probably without a bachelor's degree, (C) will be preferred.

This case study clearly points out that tree-based extensions of distance-based models successfully improve the model fitness, thereby widening their applicability. In particular, our weighted distance-based models outperform the existing distance-based ranking models.

6. Conclusion

We have in this paper investigated the extension of distance-based models using decision tree. We choose distance-based models because they are powerful tools in analyzing ranking data as they can describe the consensus situation of respondents' preference in terms of a model ranking π_0 and the degree of consensus by the parameter ϕ (or λ). However, it is obvious that the assumption of homogeneous population is sometimes unrealistic. On the other hand, decision tree can improve distance-based model by increasing the model's complexity. By incorporating tree structure in distance-based models, the assumption of homogeneous population can be relaxed. The difficulty of handling subject covariates is also overcome.

The ease of interpretation is a major strength of distance-based models. This strength can be retained when extending the model using decision tree because the decision tree itself is easy to understand.

Apart from distance-based tree models, we propose a new class of weighted distance-based models. Our weighted distance extension can keep the usual distance properties while allows the consideration of more parameters in the models and hence providing a greater flexibility for model building.

Our simulation study and case study show that our tree building algorithm can effectively split the data into different sub-populations. Tree models can handle interaction neatly and distance-based models can describe precisely the different sub-population with only a few variables. At the same time, the overall interpretation of the model is kept simple and straightforward.

Distance-based models is one out of the numerous models for ranking data, and it may not be the best of all. The tree model extension described in this paper is only one of the many possibilities. We have taken the first step here. Further efforts may include adopting the tree structure to other ranking data models, and we sincerely believe that such efforts will be meaningful and rewarding. Our C++ program source code will be available upon request.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments which greatly improved this paper. The research of Philip L.H. Yu was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. HKU 7473/05H).

References

Ahn, H., 1996. Log-normal regression modeling through recursive partitioning. Computational Statistics and Data Analysis 21, 381–398.
Audrino, F., Bühlmann, P., 2001. Tree-structured generalized autoregressive conditional heteroscedastic models. Journal of Royal Statistical Society Series B 63 (4), 727–744.
Beggs, S., Cardell, S., Hausman, J., 1981. Assessing the potential demand for electric cars. Journal of Econometrics 16, 1–19.
Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Wadsworth, Belmont, California.
Busse, L.M., Orbanz, P., Buhmann, J.M., 2007. Cluster analysis of heterogeneous rank data, in: Proceedings of the 24th International Conference on Machine Learning, pp. 113–120.
Chan, K.-Y., Loh, W.-Y., 2004. LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. Journal of Computational and Graphical Statistics 13 (4), 826–852.
Chapman, R.G., Staelin, R., 1982. Exploiting rank ordered choice set data within the stochastic utility model. Journal of Market Research 19, 288–301.
Chaudhuri, P., Lo, W.-D., Loh, W.-Y., Yang, C.-C., 1995. Generalized regression trees. Statisica Sinica 5, 641–666.
Critchlow, D.E., 1985. Metric Methods for Analyzing Partially Ranked Data. In: Lecture Notes in Statistics, vol. 34. Springer, Berlin.
Critchlow, D.E., Fligner, M.A., Verducci, J.S., 1991. Probability models on rankings. Journal of Mathematical Psychology 35, 294–318.
Diaconis, P., 1988. Group Representations in Probability and Statistics. In: Institute of Mathematical Statistics, Hayward.
Erosheva, E.A., Fienberg, S.E., Joutard, C., 2007. Describing disability through individual-level mixture models for multivariate binary data. The Annals of Applied Statistics 1 (2), 502–537.
Fligner, M.A., Verducci, J.S., 1986. Distance based ranking models. Journal of Royal Statistical Society Series B 48 (3), 359–369.
Gormley, I.C., Murphy, T.B., 2008. A mixture of experts model for rank data with applications in election studies. The Annals of Applied Statistics 2 (4), 1452–1477.
Hausman, J., Ruud, P.A., 1987. Specifying and testing econometric models for rank-ordered data. Journal of Econometrics 34, 83–104.
Inglehart, R., 1977. The Silent Revolution: Changing Values and Political Styles Among Western Publics. Princeton Univerity Press, Princeton.
Jacobs, R.A., Jorden, M.I., Nowlan, S.J., Hinton, G.E., 1991. Adaptive mixture of local experts. Neural Computation 3, 79–87.
Mallows, C.L., 1957. Non-null ranking models. I. Biometrika 44, 114–130.
Marden, J.I., 1995. Analyzing and Modeling Rank Data. Chapman and Hall.
Murphy, T.B., Martin, D., 2003. Mixtures of distance-based models for ranking data. Computational Statistics and Data Analysis 41, 645–655.
Shieh, G.S., 1998. A weighted Kendall's tau statistic. Statistics and Probability Letters 39, 17–24.
Shieh, G.S., Bai, Z., Tsai, W.-Y., 2000. Rank tests for independence—With a weighted contamination alternative. Statistica Sinica 10, 577–593.
Tarsitano, A., 2009. Comparing the effectiveness of rank correlation statistics, Working Papers, Università della Calabria, Dipartimento di Economia e Statistica, 200906.
Train, K., 2003. Discrete Choice Methods with Simulation. Cambridge University Press.
Vermunt, J.K., 2004. Multilevel latent class models. Sociological Methodology 33, 213–239.
Yu, P.L.H., 2000. Bayesian analysis of order-statistics models for ranking data. Psychometrika 65 (3), 281–299.