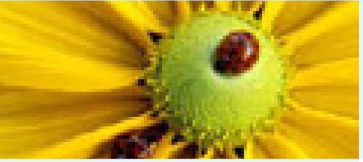
Mixtures of Weighted Distance-Based Models for Ranking Data

Paul H. Lee*
Philip L. H. Yu
The University of Hong Kong

Outline of presentation

- Introduction
- Distance-Based Models for Ranking Data
- Weighted Distance-based Models (with application)
- Simulation Studies
- Conclusions and Further Research
- Question & Answer

Introduction



Introduction

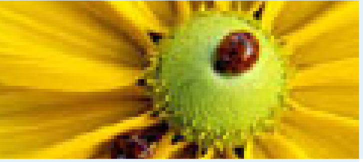
Introduction

Distance-Based Models for Ranking Data

Mixtures of Weighted Distance-based Models

Conclusions and Further Research

- What is ranking data?
  - Rank a set of items
  - Types of soft drinks
    Coke, 7-up, fanta
  - Political goals
  - Election candidates
    World footballer of the year



Introduction

- Notations used in ranking literature
  - $\boldsymbol{\pi}$: ranking
    $\pi(i)$ is the rank assigned to item i
    $\boldsymbol{\pi} = (2,4,1,3)$
    Item 1 rank 2nd, item 2 rank 4th
  - $\boldsymbol{\pi}^{-1}$: ordering
    $\pi^{-1}(i)$ is the item having rank i
    $\boldsymbol{\pi}^{-1} = (2,4,1,3)$
    Item 2 rank 1st, item 4 rank 2nd

Examples of Ranking Data

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Marketing research:
  - Green and Rao (1972): to rank 15 breakfast snack food items including toast, donut, etc.
- Travel behavior and mode of transportation:
  - Beggs, et al. (1981), Hausman, et al. (1987): to rank order 16 car designs which differed over 9 attibutes.

Examples of Ranking Data

- Politic:
  - Croon (1989): to rank 4 political goals: Order, Say, Price, and Freedom.
- Horse racing:
  - Lo et al. (1994): to predict the top two winning horses.

Types of Ranking Data

Given a set of J items. There are two types of ranking data:

- Complete rankings (rank all J items)
- Incomplete (or Partial) rankings
  - Top q rankings (select the top q items and rank them)
    When $q = 1$, top q ranking $=$ discrete choice
  - Subset rankings (select a subset of m items and rank them)
    When $m = 2$, subset ranking $=$ paired comparison
    When $m = 3$, subset ranking $=$ triple ranking

Problems of Interest

- Graphical representation of ranking data
  - visualize rankings given by judges preferably in a low-dimensional space
  - existing work: Dual scaling (Nishisato, 1994), vector models (Tucker, 1960; Carroll, 1980; Yu and Chan, 2001), ideal point models (Coombs, 1950; De Soete, et al., 1986; Yu, Chung and Leung, 2008), polyhedron representation (Thompson, 2003)

Problems of Interest

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Factor analysis
  - identify latent factors that affect ranking decision.
  - existing work: Yu, Lam and Lo (2005)
- Cluster analysis / Latent class analysis
  - find group of judges with similar rank-order preference within clusters.
  - recent work: Murphy and Martin (2003), Lee and Yu (2010)

Problems of Interest

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Modelling
  - determine probabilistic structure of probability of observing a ranking
  - existing work: a lot, see Marden (1995) for a review, Yu (2000)
  - Different types of statistical models for ranking data
    - Order-statistics
    - Paired comparison
    - Distance-based
    - Multistage
  - This talk: a weighted distance-based model?
  - mixtures models?

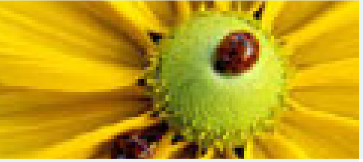
Introduction

- Properties of distance measure
  - $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_i) = 0$
  - $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = d(\boldsymbol{\pi}_j, \boldsymbol{\pi}_i)$
  - $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) > 0$ if $\boldsymbol{\pi}_i \neq \boldsymbol{\pi}_j$
- Property of metric
  Triangular inequality
  $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_k) \leq d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) + d(\boldsymbol{\pi}_j, \boldsymbol{\pi}_k)$

Distance-Based Models for Ranking Data

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Model assumption:
  - Probability of observing a ranking $\boldsymbol{\pi}$ depends on its distance to the modal ranking $\boldsymbol{\pi}_0$
  - The effect of distance is controlled by the dispersion parameter λ
- Model specification:
  - $P(\boldsymbol{\pi}|\lambda, \boldsymbol{\pi}_0) = C(\lambda)e^{-\lambda d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}$
  - $\lambda > 0$ for identification problem
  - $d(\boldsymbol{\pi}, \boldsymbol{\pi}_0)$ is the distance between $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_0$
  - $C(\lambda)$ is the proportionality constant



- Different types of distance
  - Kendall's tau
    $T(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i<j} I\{[\pi(i) - \pi(j)][\pi_0(i) - \pi_0(j)]\}$
    Used in Mallow's ϕ-model (1957)
    $P(\boldsymbol{\pi}|\phi, \boldsymbol{\pi}_0) = C(\phi)\phi^{T(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}$
  - Minimum number of pairwise adjacent transpositions needed to transform $\boldsymbol{\pi}$ to $\boldsymbol{\pi}_0$
  - Spearman's rho square
    $R^2(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_i [\pi(i) - \pi_0(i)]^2$
    Used in Mallow's θ-model (1957)
    $P(\boldsymbol{\pi}|\theta, \boldsymbol{\pi}_0) = C(\theta)\theta^{R^2(\boldsymbol{\pi}, \boldsymbol{\pi}_0)}$
    A distance but not a metric

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

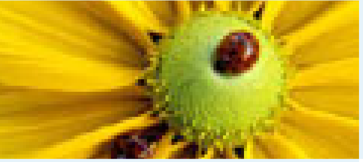
- Different types of distance
  - Spearman's rho
    $R(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \left(\sum_i [\pi(i) - \pi_0(i)]^2\right)^{0.5}$
    A metric
  - Spearman's footrule
    $F(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_i |\pi(i) - \pi_0(i)|$
- Cayley's distance
  $C(\boldsymbol{\pi}, \boldsymbol{\pi}_0) =$ minimum number of transpositions needed to transform $\boldsymbol{\pi}$ to $\boldsymbol{\pi}_0$

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Different types of distance
  - Proportionality constant $C(\lambda)$ is difficult to compute
  - Close form solution available only for:
    Kendall's tau
    Cayley's distance
  - Can be solved numerically by
    $C(\lambda) = \frac{1}{\sum_{i=1}^{k!} e^{-\lambda d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_0)}}$
- Computational time increases exponentially when number of items increase

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

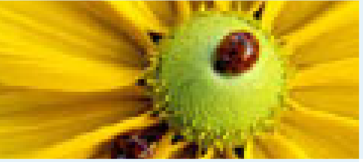
- ϕ-component model
  - Extension of Mallow's ϕ-model (Fligner and Verducci, 1988)
  - For ranking of k items, Kendall's tau can be decomposed
    $T(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_{i=1}^{k-1} V_i$
    All V's are independent
    - $V_1 = m$ means the $m + 1$st best item, with reference to $\boldsymbol{\pi}_0$, is chosen in $\boldsymbol{\pi}$
    - This item is dropped and will not be considered anymore
    - $V_2 = m$ means the $m + 1$st best item is chosen in the remaining items
    - The process is repeated until all items are ranked



- ϕ-component model
  - The V's can be weighted : $\sum_{i=1}^{k-1} \theta_i V_i$
  - The resulting model is:
    $P(\boldsymbol{\pi}|\lambda, \boldsymbol{\pi}_0) = C(\lambda) e^{-\sum_{i=1}^{k-1} \lambda_i V_i}$
    $\lambda = \{\lambda_i, i = 1, ..., k-1\}$
  - Also named $k-1$ parameter model
  - Under the re-parameterizations
    $\phi_i = e^{-\lambda_i}, i = 1, ...k-1,$
    the resulting model will be:
    $P(\pi|\phi, \pi_0) = C(\phi) \prod_{i=1}^{k-1} {\phi_i}^{V_i}$

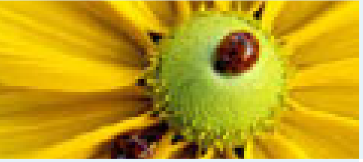
Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- The model has closed form proportionality constant if the V's are independent
- Only Kendall's tau and Cayley's distance can be decomposed in such form
- The extension based on Cayley's distance is named Cyclic structure model
- The model based on decomposition of Kendall's tau is more commonly used than Cayley's distance



Distance-Based Models for Ranking Data

- The model becomes a stage-wise process
- Properties of distance is lost
  $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) \neq d(\boldsymbol{\pi}_j, \boldsymbol{\pi}_i)$
  - $\boldsymbol{\pi}_i^{-1} = (1, 2, 3, 4), \boldsymbol{\pi}_j^{-1} = (2, 3, 4, 1)$
    $V_1 = 3, V_2 = 0, V_3 = 0$
  - $\boldsymbol{\pi}_i^{-1} = (2, 3, 4, 1), \boldsymbol{\pi}_j^{-1} = (1, 2, 3, 4)$
    $V_1 = 1, V_2 = 1, V_3 = 1$
  - In general, $3\lambda_1 + 0\lambda_2 + 0\lambda_3 \neq \lambda_1 + \lambda_2 + \lambda_3$
- Find an extension which
  - Retains the properties of distance
  - Allows weights for different rank



- Weighted distance
- Inspired by Shieh (1998, 2000)
- Different weights for different rank, according to π_0
  - Weighted Kendall's tau
    $T_w(\boldsymbol{\pi}, \boldsymbol{\pi}_0) =$
    $\sum_{i<j} w_{\pi_0(i)} w_{\pi_0(j)} I\{[\pi(i) - \pi(j)][\pi_0(i) - \pi_0(j)]\}$
  - Weighted Spearman's rho square
    $R_w^2(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_i w_{\pi_0(i)} [\pi(i) - \pi_0(i)]^2$
  - Weighted Spearman's rho
    $R_w(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \left(\sum_i w_{\pi_0(i)} [\pi(i) - \pi_0(i)]^2\right)^{0.5}$
  - Weighted Spearman's footrule
    $F_w(\boldsymbol{\pi}, \boldsymbol{\pi}_0) = \sum_i w_{\pi_0(i)} |\pi(i) - \pi_0(i)|$

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Properties of distance is retained
  $d(\boldsymbol{\pi}_i, \boldsymbol{\pi}_j) = d(\boldsymbol{\pi}_j, \boldsymbol{\pi}_i)$

- Example : Spearman's rho square
  Let $R_a = [\pi_i(a) - \pi_j(a)]^2$

  - $\boldsymbol{\pi}_i^{-1} = (1, 2, 3, 4), \boldsymbol{\pi}_j^{-1} = (2, 3, 4, 1)$
    $R_1 = 9, R_2 = 1, R_3 = 1, R_4 = 1$
  - $\boldsymbol{\pi}_i^{-1} = (2, 3, 4, 1), \boldsymbol{\pi}_j^{-1} = (1, 2, 3, 4)$
    $R_1 = 9, R_2 = 1, R_3 = 1, R_4 = 1$
  - In general, $w_2 + w_3 + w_4 + 9w_1 = w_2 + w_3 + w_4 + 9w_1$
  - Note : before swapping, w_1 : weight for item ranked first in $\boldsymbol{\pi}_j$
    After swapping, w_1 : weight for item ranked first in $\boldsymbol{\pi}_i$

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

Mixtures of Weighted Distance-based Models

Mixtures of Weighted Distance-based Models

- Distance-based models assume single modal ranking $\boldsymbol{\pi}_0$
- Relax this assumption using mixtures models
- Probability of observing a ranking $\boldsymbol{\pi}$ from a mixtures of G weighted distance-based models:

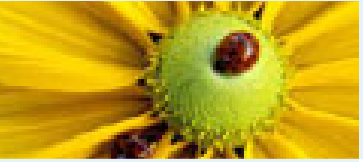
  $$P(\boldsymbol{\pi}) = \sum_{g=1}^{G} p_g P(\boldsymbol{\pi}|\mathbf{w}_g, \boldsymbol{\pi}_{0g}) = \sum_{g=1}^{G} p_g \frac{e^{-d_{\mathbf{w}_g}(\boldsymbol{\pi},\boldsymbol{\pi}_{0g})}}{C(\mathbf{w}_g)}$$

  - p_g is the proportion of observations belong to group g
  - $\mathbf{w}_g$, $\boldsymbol{\pi}_{0g}$ are the model parameters of group g

Mixtures of Weighted Distance-based Models

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Use EM algorithm to obtain MLE
  - E-step: for all observations, compute the probabilities of belonging to every sub-population
  - M-step: maximize the conditional expected complete-data loglikelihood
- Use BIC ($-2\ell + v\log(n)$) to determine the number of mixtures
  - ℓ is the loglikelihood
    $\ell = \sum_{i=1}^{n} \log\left(\sum_{g=1}^{G} p_g \frac{e^{-d_{\mathbf{w}_g}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{0g})}}{C(\mathbf{w}_g)}\right)$
  - v is the number of parameters
  - n is the number of observations



- EM algorithm:
  - Define $z_i = (z_{1i}, ..., z_{Gi})$: $z_{gi} = 1$ if $i \in g$, otherwise $z_{gi} = 0$
  - Complete loglikelihood:
    $L_{com} = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{gi}[\log(p_g) - d_{\mathbf{w}_g}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{0g}) - log(C(\mathbf{w}_g))]$
  - E-step: compute $\hat{z}_{gi}$ by:
    $\hat{z}_{gi} = \frac{\hat{p}_g P(\hat{\boldsymbol{\pi}}_i | \hat{\mathbf{w}}_g, \hat{\boldsymbol{\pi}}_{0g})}{\sum_{h=1}^{G} \hat{p}_h P(\hat{\boldsymbol{\pi}}_i | \hat{\mathbf{w}}_h, \hat{\boldsymbol{\pi}}_{0h})}$
  - M-step compute $\hat{\mathbf{w}}_g$ and $\hat{\boldsymbol{\pi}}_{0g}$ by solving:
    $\frac{\sum_{i=1}^{n} \hat{z}_{gi} d_{\mathbf{w}_g}(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{0g})}{\sum_{i=1}^{n} \hat{z}_{gi}} = \sum_{j=1}^{k!} P(\boldsymbol{\pi}_j | \mathbf{w}_g, \boldsymbol{\pi}_{0g}) d_{\mathbf{w}_g}(\boldsymbol{\pi}_j, \boldsymbol{\pi}_{0g})$

Simulation Studies

- Two simulation studies
- Aims of the two studies:
  1. Performance of estimation algorithm
  2. Effectiveness of BIC

Simulation Studies

- Ranking of 4 items, with 2000 observations
- Generate 50 times
- Simulation settings:

Model	π_0	w_1	w_2	w_3	w_4
1	$1 \succ 2 \succ 3 \succ 4$	2	1.5	1	0.5
2	$1 \succ 2 \succ 3 \succ 4$	1	0.75	0.5	0.25

Model	p	π_0	w_1	w_2	w_3	w_4
3	0.5	$1 \succ 2 \succ 3 \succ 4$	2	1.5	1	0.5
	0.5	$4 \succ 3 \succ 2 \succ 1$	2	1.5	1	0.5
4	0.5	$1 \succ 2 \succ 3 \succ 4$	2	1.5	1	0.5
	0.5	$4 \succ 3 \succ 2 \succ 1$	1	0.75	0.5	0.25

Simulation Studies 1

- Compute MLE, assume number of mixtures is given
- Parameter estimates:

	Model 1	Model 2
π_0	$1 \succ 2 \succ 3 \succ 4$	$1 \succ 2 \succ 3 \succ 4$
w_1	2.002(0.059)	0.981(0.081)
w_2	1.509(0.055)	0.779(0.089)
w_3	0.995(0.032)	0.492(0.035)
w_4	0.497(0.013)	0.250(0.030)

Simulation Studies 1

Introduction

Distance-Based Models for Ranking Data

Mixtures of Weighted Distance-based Models

Conclusions and Further Research

- Results:

	Model 3		Model 4	
$\boldsymbol{\pi}_0$	$1 \succ 2 \succ 3 \succ 4$	$4 \succ 3 \succ 2 \succ 1$	$1 \succ 2 \succ 3 \succ 4$	$4 \succ 3 \succ 2 \succ 1$
p	0.500(0.007)	0.500	0.499(0.028)	0.501
w_1	1.976(0.129)	1.961(0.123)	2.088(0.232)	1.039(0.158)
w_2	1.535(0.121)	1.540(0.107)	1.458(0.173)	0.747(0.174)
w_3	0.995(0.063)	0.995(0.065)	1.036(0.182)	0.497(0.072)
w_4	0.500(0.035)	0.498(0.025)	0.501(0.050)	0.252(0.072)

- Estimation method is accurate
- Accuracy increases for larger w

Simulation Studies 2

- Use BIC to select the number of mixtures
- Selection frequencies:

Model	N	1	$1+N$	2	$2+N$	3
1	0	45	5	0	0	0
2	0	37	13	0	0	0
3	0	0	0	49	1	0
4	0	0	0	47	3	0

- BIC can identify the number of mixtures most of the time
- BIC sometimes suggest including an additional noise component ($\mathbf{w}=0$)

Application on Real data

- Dataset description:
  - Political studies from Croon (1989)
  - 2262 respondents from Germany
  - Rankings of 4 political goals

Application on Real data

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Dataset description:
  - Respondents ranked 4 political goals for their Government
    (A) Maintain order in nation
    (B) Give people more to say in Government decisions
    (C) Fight rising prices
    (D) Protect freedom of speech
  - Respondents can be classified:
    "Materialist" : top 2 = (A) and (C)
    "Post-materialist" : top 2 = (B) and (D)
    "Mixed" : other combinations

Application on Real data

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Best model: F_w, 3 groups of mixture
- BIC: 12670.82
- Better than Strict Utility model (12670.87) and Pendergrass-Bradley model (12673.07) in Croon (1989)

Group	Ordering	p	w_1	w_2	w_3	w_4
1	$C \succ A \succ B \succ D$	0.352	2.030	1.234	~ 0	0.191
2	$A \succ C \succ B \succ D$	0.441	1.348	0.917	0.107	0.104
3	$B \succ D \succ C \succ A$	0.208	0.314	~ 0	0.151	0.552

Application on Real data

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Groups 1 and 2: Materialists
  Items (A) and (C) are preferred
  w_1 and w_2 are large, positions of (A) and (C) are stable

- Group 3: Post-materialists
  Items (B) and (D) are preferred
  all weights are small, positions of items are not stable

Conclusions and Further Research

Conclusions and Further Research

Introduction
Distance-Based Models for Ranking Data
Mixtures of Weighted Distance-based Models
Conclusions and Further Research

- Conclusions
  - Flexibility increased
  - Assumption of homogeneous population is relaxed