A Generative Model for Rank Data Based on an Insertion Sorting Algorithm

J. Jacques & C. Biernacki

Laboratory of Mathematics, UMR CNRS 8524 & University Lille 1 (France)

COMPSTAT'2010

Outline

Ranking and ordering notations

Objects to rank

Three holidays destinations:

$$\mathcal{O}_1 = \text{Campaign},\ \mathcal{O}_2 = \text{Mountain and } \mathcal{O}_3 = \text{Sea}$$

Rank notations

- Unformalized: First Sea, second Campaign, and last Mountain
- Ordering:

$$x = (3, 1, 2) = (\overset{1^{\text{st}}}{\mathcal{O}_3}, \overset{2^{\text{nd}}}{\mathcal{O}_1}, \overset{3^{\text{th}}}{\mathcal{O}_2})$$

- Ranking:

$$x^{-1} = (2, 3, 1) = (\overset{\mathcal{O}_1}{2^{\text{nd}}}, \overset{\mathcal{O}_2}{3^{\text{th}}}, \overset{\mathcal{O}_3}{1^{\text{st}}})$$

Interest of rank data

Human activities involving preferences, attitudes or choices

Web Page ranking	Sport
Sociology	Politics
Economics	Educational Testing
Biology	Psychology
Marketing	...

They often result from a transformation of other kinds of data!

A model of reference: Mallows Φ ($\sim$1950)

$$\mathrm{pr}(x;\mu,\theta) \propto \exp(-\theta d_K(x,\mu))$$

- $\mu = (\mu_1, \ldots, \mu_m)$: Rank of reference parameter (m objects)
- $d_K(x,\mu)$: Kendall distance between $x = (x_1, \ldots, x_m)$ and μ
- $\theta \in \mathbb{R}^+$: Dispersion parameter
  - $\theta > 0$: μ is the mode and dispersion decreases with θ
  - $\theta = 0$: Uniformity (max. of dispersion)

Interesting ...

- Many other models are linked with it
- Other distances can be retained (Cayley...)

Motivation for an alternative model

Two fundamental hypotheses

1. x results from a sorting algo. based on paired comparisons
2. $\neq$ between x and μ only result from bad paired comparisons

$\Rightarrow$ Mallows Φ model can be interpreted as a sorting algorithm where all pairs comparisons are performed.

$\Downarrow$

Minimizing errors $\Leftrightarrow$ minimizing paired comparisons

If $m \leq 10$, the insertion sorting algorithm has to be retained

$\Downarrow$

The present work!

Formalize, study, estimate and experiment a new model...

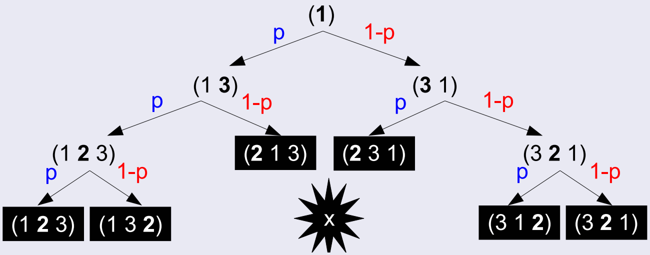
Outline

1. Motivation
   - Importance of rank data
   - Models for rank data
2. The Insertion Sorting Rank model
   - Formalization
   - Properties
   - Estimation of the model parameters
3. Numerical illustration
   - Comparison of ISR and Mallows Φ
   - A specificity of ISR: Initial rank σ
4. Concluding remarks

Notations

- $x = (x_1, \ldots, x_m)$: Observed rank
- $\mu = (\mu_1, \ldots, \mu_m)$: Rank of reference parameter ("true" rank)
- $p \in [0, 1]$: Probability of good paired comparison (parameter)
- $\sigma = (\sigma_1, \ldots, \sigma_m)$: Initial rank (latent data!)

Example: $\mu = (1, 2, 3)$ and $\sigma = (1, 3, 2)$



Model expression

- good(x, σ, μ): Total number of good paired comparisons
- bad(x, σ, μ): Total number of bad paired comparisons

$$\mathrm{pr}(x|\sigma; \mu, p) = p^{\mathrm{good}(x,\sigma,\mu)} \, (1-p)^{\mathrm{bad}(x,\sigma,\mu)}$$

But σ is latent: Marginal over $p(\sigma) = m!^{-1}$

$$\mathrm{pr}(x; \mu, p) = m!^{-1} \sum_{\sigma} \mathrm{pr}(x|\sigma; \mu, p)$$

Properties of the ISR model

Well-behaved model

- μ is the mode and $\bar{\mu}$ the anti-mode ($p > \frac{1}{2}$)
- $\mathrm{pr}(\mu; \mu, p) - \mathrm{pr}(x; \mu, p)$ is an increasing function of p
- Identifiability of (μ, p) if $p > \frac{1}{2}$
- Uniform distribution when $p = \frac{1}{2}$

Space reduction for p

Symmetry: $\mathrm{pr}(x; \bar{\mu}, 1 - p) = \mathrm{pr}(x; \mu, p) \quad \Rightarrow \quad p \in [\frac{1}{2}, 1]$

The EM algorithm

Maximizing the likelihood from incomplete data $(x^1, \ldots, x^n)$

- E step:

$$t_{i\sigma} = \mathrm{pr}(\sigma | x^i; \mu, p) = \frac{\mathrm{pr}(x^i | \sigma; (\mu, p))}{\sum_s \mathrm{pr}(x^i | s; (\mu, p))}$$

- M step: μ^+ given by browsing the half space (symmetry)

$$p^+ = \frac{\sum_{i=1}^n \sum_\sigma t_{i\sigma} \mathrm{good}(x^i, \sigma, \mu)}{\sum_{i=1}^n \sum_\sigma t_{i\sigma} (\mathrm{good}(x^i, \sigma, \mu) + \mathrm{bad}(x^i, \sigma, \mu))}$$

Possibility to restrict the candidates μ...

...to a stochastic subset of $(x^1, \ldots, x^n)$ related to empirical freq.

Outline

Five real data sets

Data set	Quizz	m	n	μ^*	Objects $\mathcal{O}_1, \ldots, \mathcal{O}_m$
Rank the four national football teams according to increasing number of victories in the football World Cup					
Football	Yes	4	40	(1,2,4,3)	France, Germany, Brasil, Italy
Rank chronologically these Quentin Tarantino movies					
Cinema	Yes	4	40	(3,2,4,1)	Inglourious Basterds, Pulp Fiction Reservoir Dogs, Jackie Brown
Results of the four nations rugby league, from 1910 to 1999 (except years where they were tie)					
Rugby 4N	No	4	20	None	England, Scotland, Ireland, Walles
Rank five words according to strength of association (least to most associated) with the target word "Idea"					
Word association	Yes	5	98	None	Thought, Play, Theory, Dream, Attention
Rank seven sports according to their preference in participating					
Sports	Yes	7	130	None	Baseball, Football, Basketball, Tennis, Cycling, Swimming, Jogging

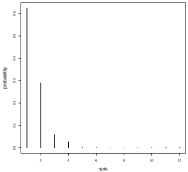
Results

Data set	Model	$\hat{\mu}$	$\hat{p}$ / $\hat{\theta}$	L	$\widehat{\text{p-value}}$	$\#\mu$	Time (s)
Football	ISR	(1,2,4,3)	0.834	-89.58	0.001	1	1.6
	Φ	(1,2,4,3)	1.093	-90.22	0.001	1	3.0
Cinema	ISR	(4,3,2,1)	0.723	-112.99	0.042	14	4.2
	Φ	(4,3,2,1)	0.627	-113.16	0.029	2	7.3
Rugby 4N	ISR	(2,4,1,3)	0.681	-59.53	0.538	12	2.7
	Φ	(2,4,1,3)	0.528	-59.18	0.395	2	7.0
Word association	ISR	(2,5,4,3,1)	0.879	-283.00	0.001	1	6.0
	Φ	(2,5,4,3,1)	1.432	-252.57	0.019	1	19.0
Sports	ISR	(1,3,2,4,5,7,6)	0.564	-1103.50	0.999	1	1353.1
	Φ	(1,3,4,2,5,6,7)	0.080	-1104.24	0.045	11	15842

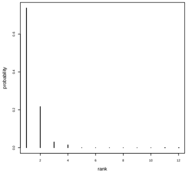
- Both models are hard competitors
- Computational feasibility, even for $m = 7$
- Efficiency of μ space restriction (both models)
- Consistency in the $\hat{p}/\hat{\theta}$ meaning: $\hat{p}_{\text{football}} > \hat{p}_{\text{cinema}}$ and $\hat{\theta}_{\text{football}} > \hat{\theta}_{\text{cinema}}$
- Often both models with same $\hat{\mu}$ except "Sports": ISR more coherent?
- Parameter p of ISR easier to understand

ISR detects quizz or no-quizz through $\hat{\sigma}$!

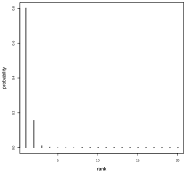
$$\mathrm{pr}(\sigma^1 = \ldots = \sigma^n = s | x^1, \ldots, x^n, \sigma^1 = \ldots = \sigma^n; \hat{\mu}, \hat{p})$$



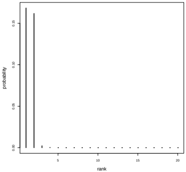
Football



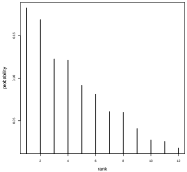
Cinema



Word



Sports



Rugby 4N (no-quizz!)

Outline

1. Motivation
   - Importance of rank data
   - Models for rank data
2. The Insertion Sorting Rank model
   - Formalization
   - Properties
   - Estimation of the model parameters
3. Numerical illustration
   - Comparison of ISR and Mallows Φ
   - A specificity of ISR: Initial rank σ
4. Concluding remarks

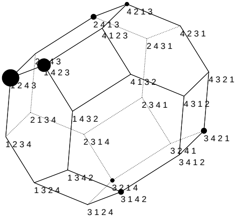
Summary about the ISR proposal

- Optimality when $m \leq 10$: Minimize number of errors
- Meaningful parameters
- The initial rank σ is taken into account and meaningful
- Good results when compare to the Mallows Φ
- Computational feasible for $m \leq 7$ in R, probably 10 with C
- Estimation easy with an EM algorithm
- Efficient starting strategy for avoiding combinatory about μ

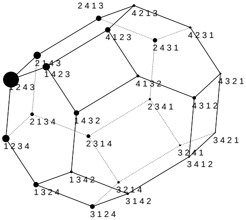
Future work

- $m \leq 10$: Try non-optimal but realistic sorting algorithms
- $m > 10$: Which sorting algorithm? Computational cost?

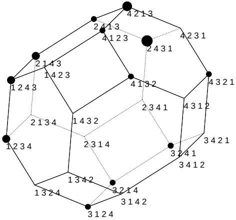
Polytopes illustration



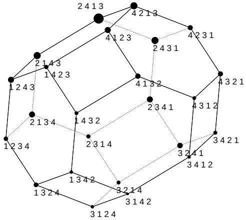
Empirical “Football”



estimate ISR “Football”



Empirical “Rugby 4N”



estimate ISR “Rugby 4N”

Application to clustering of rank data

Natural extension to clustering by assuming that observed ranks arise from a **mixture** of K ISR distributions

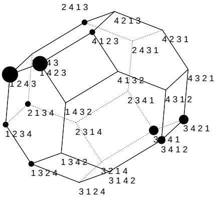
$$\mathrm{pr}(x;\theta) = \sum_{k=1}^{K} \frac{\pi_k}{m!} \sum_{\sigma} \mathrm{pr}(x|\sigma;\mu_k,p_k)$$

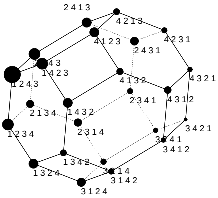
where

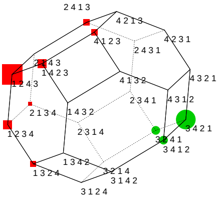
- $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, p_1, \ldots, p_K)$
- $\mathrm{pr}(x|\sigma;\mu_k,p_k) = p_k^{\mathrm{good}(x,\sigma,\mu_k)} (1-p_k)^{\mathrm{bad}(x,\sigma,\mu_K)}$

An example : Football Quizz

Rank these teams in increasing order of victories number to the Football World Cup : 1. France 2. Germany 3. Brasil 4. Italy







empiric | ISR | Mixture ISR

	ISR	Mixture ISR	
μ	$(1,2,4,3)$	$(1,2,4,3)$	$(3,4,2,1)$
p	0.69	0.85	0.84
π		0.73	0.27
BIC	179.1	**160.6**	