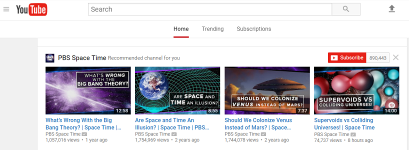
Solving Large Linear-ordering Problems

R. Kohli [1] – Kh. Boughanmi [1] – V. Kohli [2]

[1] Columbia Business School

[2] Northwestern University

The Importance of Ranking



1. Youtube videos
2. Universities ranking
3. Social networks feed

. . .

World Rank	Institution*	Country /Region	National Rank	Total Score	Score on Alumni
1	Harvard University		1	100	100
2	Stanford University		2	72.1	41.8
3	Massachusetts Institute of Technology (MIT)		3	70.5	68.4
4	University of California-Berkeley		4	70.1	66.8
5	University of Cambridge		1	69.2	79.1
6	Princeton University		5	60.7	52.1
7	California Institute of Technology		6	60.5	48.5
8	Columbia University		7	59.6	65.1
9	University of Chicago		8	57.4	61.4
9	University of Oxford		2	57.4	51

Presentation Structure

Problem Structure

- Statistical Approach
- Discrete Optimization Approach

Relation Between The Two Methods

- Randomized Algorithm

Application: Ordering Funny Youtube Videos

Problem Structure

Problem Structure

Take a set X of m objects

- **Data:**
  - Paired comparisons (not all pairs) over objects
  - Partial or complete rankings
- **Output:**
  - Rank ordering of all objects in X that best represent the data

Statistical Approach

- Let $u_i = v_i + \epsilon_i$ denote the utility of item i in X.
- Rank ordered logit

$$P(u_1 > u_2 > \cdots > u_m) = \frac{e^{v_1}}{\sum_{j=1}^{m} e^{v_j}} \cdot \frac{e^{v_2}}{\sum_{j=2}^{m} e^{v_j}} \cdots \frac{e^{v_{m-1}}}{\sum_{j=m-1}^{m} e^{v_j}}$$

- Order by $v_1, v_2, \ldots, v_n$ if $v_1 > v_2 > \cdots > v_n$.
- Properties:
    - Random utility model
    - Covariates can be added
    - Partial ranking data or paired comparison data
    - Fast (convex optimization problem)

Discrete Optimization Approach (Kemeny 1959)

- Data: Paired comparisons
- Objective: Find a single ordering of elements $x \in X$ such that the number of non-reversal between predicted and actual is maximized
- Properties:
    - NP-Hard
    - $O(m^2)$ decision variables
    - $O(m^3)$ constraints
    - Many approximation algorithms: Grötschel, Jünger and Reinelt (1984), Laguna et al (1999), Schiavinotto and Stützle(2004), García et al. (2006), Charon and Hudry (2007), Campos et al. (2001), Ailon et al. (2008), Kenyon-Mathieu and Schudy (2007), Van Zuylen and Williamson (2009), Martí and Reinelt (2011), Charon and Hudry (2007, 2010), Fagin et al. (2006), Filkov and Skiena (2004), Van Zuylen and Williamson (2009)...

Discrete Optimization Approach (Kemeny 1959)

- Let $x_{ij} = 1$ if alternative i precedes alternative j in a linear ordering; otherwise, $x_{ij} = 0$
- Let n_{ij} the number of times alternative i beats alternative j
- The optimal linear ordering z^* is a solution to the following 0-1 integer programming problem:

$$\text{Maximize} \;\; z = \sum_{i=1}^{m} \sum_{j=1, j \neq i}^{m} n_{ij} x_{ij}.$$

$$\text{subject to} \;\; x_{ij} + x_{ji} = 1, \text{ for all } \; i \neq j, 1 \leq i, j \leq m,$$

$$x_{ij} + x_{jk} + x_{ki} \leq 2, \text{ for all } i \neq j \neq k, \; 1 \leq i, j, k \leq m,$$

$$x_{ij} \in \{0, 1\}, \text{for all } i \neq j, \; 1 \leq i, j \leq m.$$

Relation Between The Two Methods

- Is one method better than the other?
- Do they produce similar aspects?
- How are the two related?

Relation Between The Two Methods

Randomized Algorithm

Let:

- $u_i = v_i + \epsilon_i$ denote the random utility of alternative i
- ϵ_i has an independent, extreme value distribution, for each $i = 1, \ldots, m$

Suppose we knew the values of $v_1, \ldots, v_m$. Then we could use the following randomized algorithm to obtain a linear ordering:

(1) Generate an observation $u_i = v_i + \epsilon_i$

(2) Arrange the utilities u_i in decreasing order of their values

Continuous Formulation

The expected value of the solution obtained by the randomized algorithm is

$$E = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \left(\frac{e^{v_i}}{e^{v_i} + e^{v_j}} \right) n_{ij} + \left(\frac{e^{v_j}}{e^{v_i} + e^{v_j}} \right) n_{ji},$$

where

- n_{ij} is the number of pairs in which i is preferred to j
- n_{ji} is the number of pairs in which j is preferred to i

Let E^* denote the value of the optimal solution to the problem of maximizing E over the variables $v_1, \ldots, v_m$.

Continuous/Discrete Formulation Equivalence

Theorem

$E^* = z^*$

Implications:

- Transformed a linear constrained 0-1 integer program into a continuous unconstrained non-linear program

Relation Between The Two Approaches

Method	Maximum likelihood	Maximum non-reversals	Maximum expected value
Objective	$\max L = \prod_{i \neq j} p_{ij}^{n_{ij}}$	$\max z = \sum_{i \neq j} n_{ij} x_{ij}$	$\max E = \sum_{i \neq j} n_{ij} p_{ij}$
Parameters	$m - 1$	$\frac{m(m-1)}{2}$	$m - 1$
Constrains	–	$O(m^3)$	–

Geometric mean $\leq$ Arithmetic mean

$$L^{1/n} \leq \frac{E}{n}$$

$$\Downarrow$$

$$\frac{1}{2} \leq L^{*\ 1/n} \leq \frac{E^*}{n}$$

Lower Bound On The Performance Ratio

Let

$$M = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \max\{n_{ij}, n_{ji}\}$$

then if $r = \frac{E}{E^*}$,

$$r \geq k^* \left[L^{*1/n} + \max\left\{ \frac{1}{N}\left(\sqrt{\hat{p}_N} - \sqrt{\hat{p}_1}\right)\right)^2, \mathrm{Var}(\sqrt{\hat{p}}) \right\} \right],$$

where $k^* = N/z^* > N/M \geq 1$ (Tung (1975) and Aldaz (2012)).

Application: Ordering Funny Youtube Videos

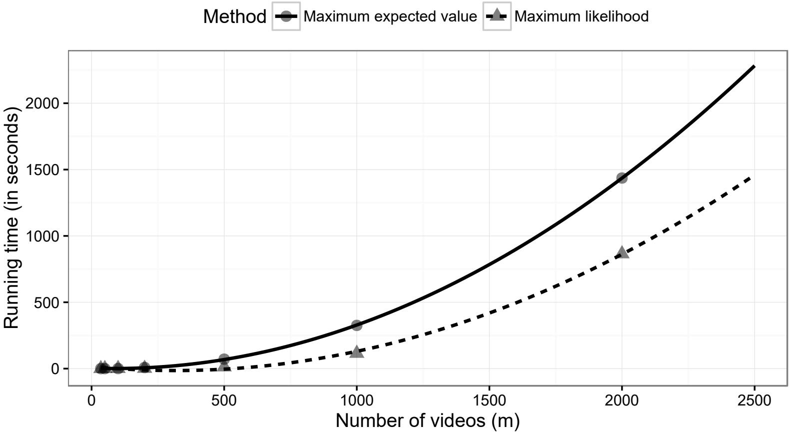
Comedy Slam

Comedy Slam
Subscribe
Which one's funnier?
Man Punches Hole Through Computer P...
21D242A3
7:40:33 PM
MAY 21 2006
0:00 / 0:37
VS
Skip
Dad scaring baby by snoring
0:00 / 0:32
Vote for me
Vote for me

Summary of Data for The Solved Problems

No. of videos	No. of pairs of videos	No. of paired comparisons	% of videos	% pairs of videos	% of paired comparisons
35	595	359,326	0.17	0.18	31.56
50	1,138	396,033	0.24	0.35	34.78
100	2,844	462,589	0.47	0.87	40.63
200	6,709	549,131	0.94	2.05	48.23
500	20,581	594,723	2.36	6.29	52.23
1,000	43,554	653,455	4.72	13.32	57.39
2,000	75,872	715,166	9.43	23.20	62.81

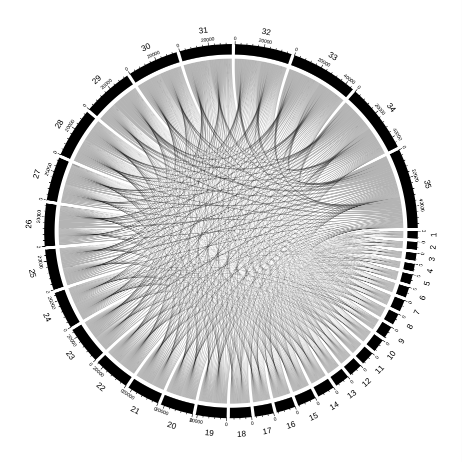
Computational times as a function of the number of alternatives in a problem.



Solution values and lower bounds on the performance ratios

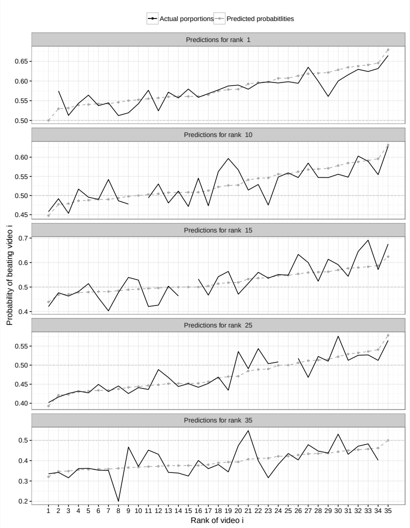
No. of videos	z_ℓ/M	z_E/M	L.B. on r	Kendall's τ
35	0.9952	0.9972	0.9067	0.8454
50	0.9921	0.9951	0.9057	0.8057
100	0.9893	0.992	0.9029	0.6008
200	0.9853	0.9883	0.8955	0.5427
500	0.9686	0.9731	0.8787	0.5670
1,000	0.9467	0.9529	0.8578	0.5506
2,000	0.9236	0.9319	0.8339	0.4877

Paired comparisons data for the problem with 35 videos

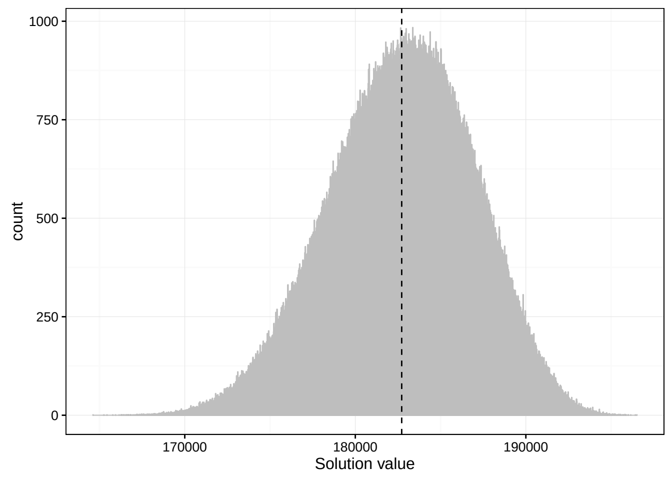


Note: Black bands correspond to videos and edges to a paired comparisons.

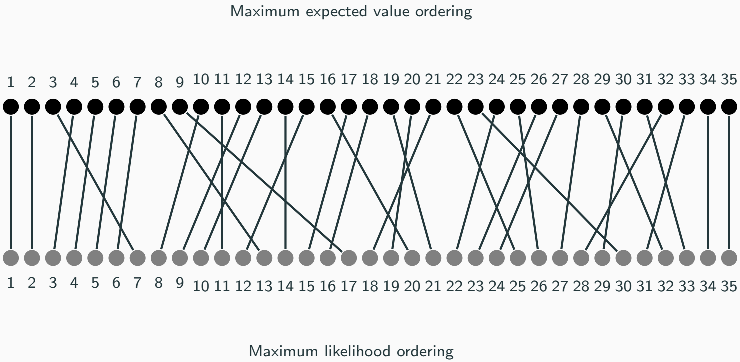
Actual and predicted proportions of votes for the kth ranked video against the other thirty-four videos.



Distribution of solution values for the randomized algorithm using the maximum likelihood solution



Comparison of the obtained rank orderings



Conclusion

- We examined the relationship between continuous and discrete optimization
- We introduced a randomized algorithm to solve large ordering problems
- We illustrated the different approaches on ranking Youtube videos

Questions?

Thank you