

A GEOMETRICAL EXTENSION OF THE BRADLEY-TERRY MODEL

INFORMATION GEOMETRY OF RANKING PROBLEM

Noboru Murata

June 20, 2023

<https://noboru-murata.github.io/>

Introduction

- Bradley-Terry model

- conventional estimation algorithm

Problem Formulation

- geometrical overview

Illustrative Example

- reguralization property

- weight adaptation with local influence

- grouped ranking data

Conclusion

INTRODUCTION

Win-Loss Standings of MLB (American East)

	Yankees	Rays	Red Sox	Blue Jays	Orioles
Yankees	-	6	8	9	5
Rays	8	-	7	8	7
Red Sox	6	9	-	8	9
Blue Jays	5	4	4	-	?
Orioles	7	8	5	?	-

Win-Loss Standings of MLB (American East)

	Yankees	Rays	Red Sox	Blue Jays	Orioles
Yankees	-	6	8	9	5
Rays	8	-	7	8	7
Red Sox	6	9	-	8	9
Blue Jays	5	4	4	-	?
Orioles	7	8	5	?	-

Problem

- estimate intrinsic strengths of teams
- predict results of unobserved matches

notations:

- i : a member of k individuals
(e.g. *baseball team*)
- θ_i : skill of individual i
(e.g. *strength of team*)
- probability model (binomial distribution):

$$\Pr\{i \text{ beats } j\} = \Pr(i \succ j) = \frac{\theta_i}{\theta_i + \theta_j},$$

(e.g. *win-loss probability between teams i and j*)

- n_{ij} : observation, i.e. the number of times that i beats j
(Bradley and Terry 1952)

- two sets of binomial distributions

- data: $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$

$$P_{\mathcal{D}_{ij}}^{(b)}(i \succ j) = \frac{n_{ij}}{n_{ij} + n_{ji}}$$

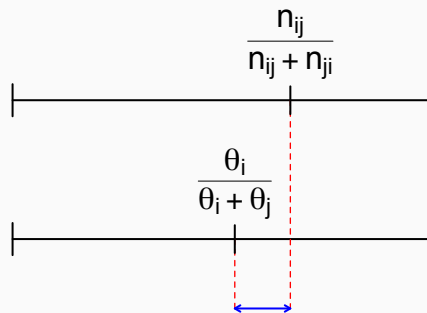
- model: $\theta_{ij} = \{\theta_i, \theta_j\}$

$$P_{\theta_{ij}}^{(b)}(i \succ j) = \frac{\theta_i}{\theta_i + \theta_j}$$

- compare distributions

- discrepancy (KL divergence):

$$\text{Dist}(\{n_{ij}, n_{ji}\}, \{\theta_i, \theta_j\}) = D(P_{\mathcal{D}_{ij}}^{(b)}, P_{\theta_{ij}}^{(b)})$$



$$\text{Dist}(\{n_{ij}, n_{ji}\}, \{\theta_i, \theta_j\})$$

conventional algorithm (Hastie and Tibshirani 1998)

- objectives: likelihood of binomial distribution

$$\begin{aligned} L(\theta) &= - \sum_{i < j} \left(n_{ij} \log \frac{\theta_i}{\theta_i + \theta_j} + n_{ji} \log \frac{\theta_j}{\theta_i + \theta_j} \right) \\ &= \sum_{i < j} (n_{ij} + n_{ji}) D(P_{\mathcal{D}_{ij}}^{(b)}, P_{\theta_{ij}}^{(b)}) + \text{const.} \end{aligned}$$

- iterative updates:

- calculate:

$$\theta_i \leftarrow \frac{\sum_{j \neq i} n_{ij}}{\sum_{j \neq i} \frac{n_{ij} + n_{ji}}{\theta_i + \theta_j}}$$

- re-normalize: $\|\theta\|_1 = 1$

PROBLEM FORMULATION

basic ideas (Fujimoto, Hino, and Murata 2011)

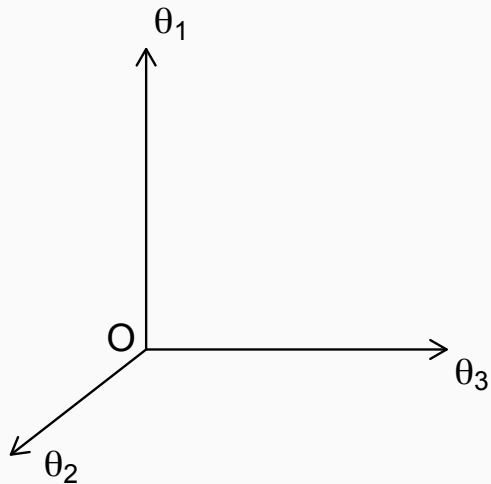
- BT model parameter can be identified with a multinomial distribution
- pairwise comparison data can be regarded as incomplete data from multinomial distributions

basic ideas (Fujimoto, Hino, and Murata 2011)

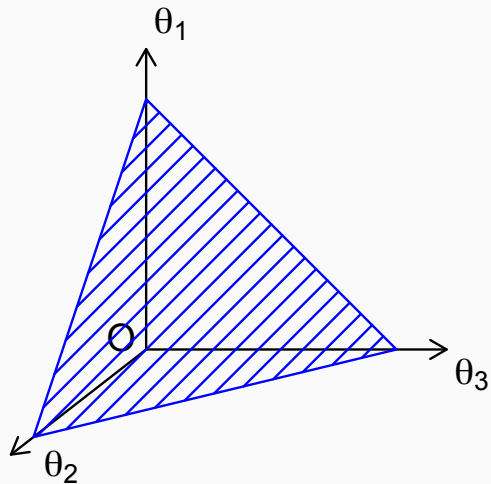
- BT model parameter can be identified with a multinomial distribution
a point on the probability simplex
- pairwise comparison data can be regarded as incomplete data from
multinomial distributions
an m -flat manifold in the probability simplex

em-Algorithm (Amari, 1995)

optimal parameter can be obtained by means of iterative e and m -projections

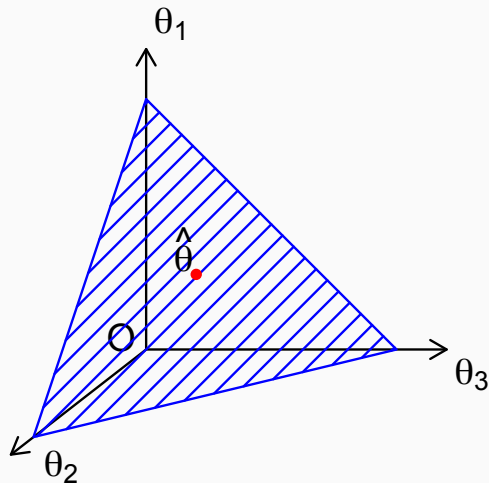


example: $k = 3$



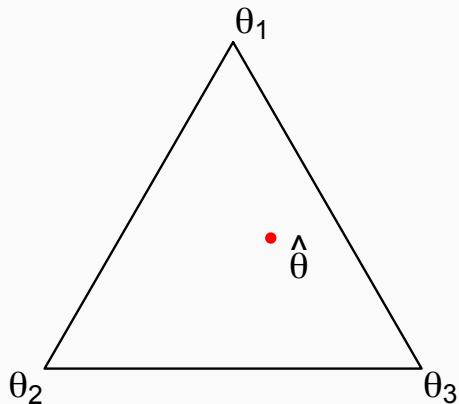
example: $k = 3$

- $\theta_i \geq 0$ (positivity)
- $\sum \theta_i = 1$ (normalized)

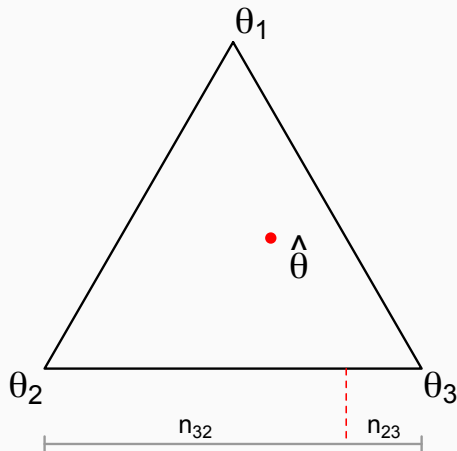


example: $k = 3$

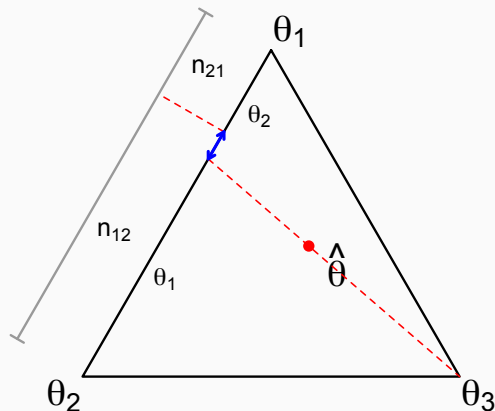
- $\theta_i \geq 0$ (positivity)
- $\sum \theta_i = 1$ (normalized)
- estimate $\hat{\theta}$
(a point in the simplex)



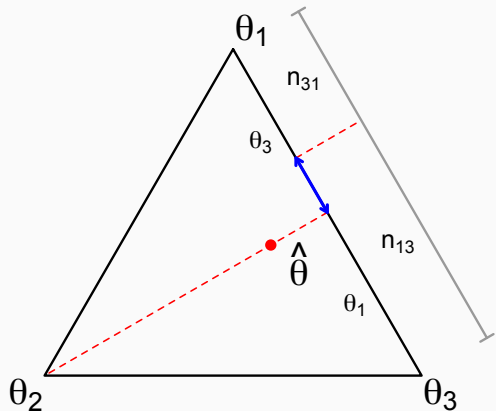
- $\hat{\theta}$: current estimate



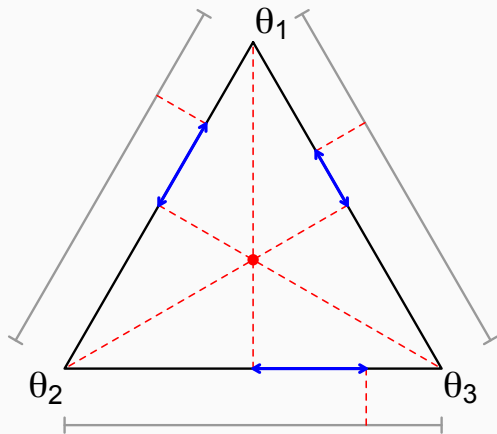
- $\hat{\theta}$: current estimate
- construct $P_{\mathcal{D}_{ij}}^{(b)}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$



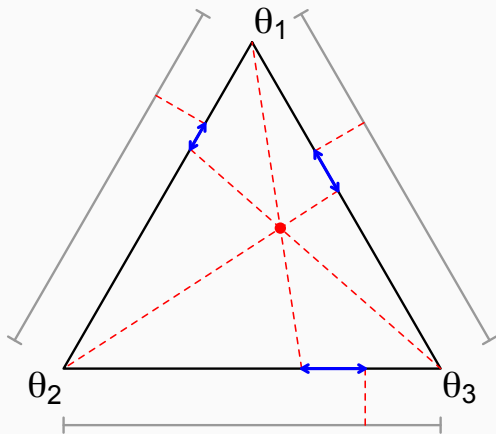
- $\hat{\theta}$: current estimate
- construct $P_{\mathcal{D}_{ij}}^{(b)}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$
- construct $P_{\theta_{ij}}^{(b)}$ from $\theta_{ij} = \{\theta_i, \theta_j\}$
- compare all the possible pairs



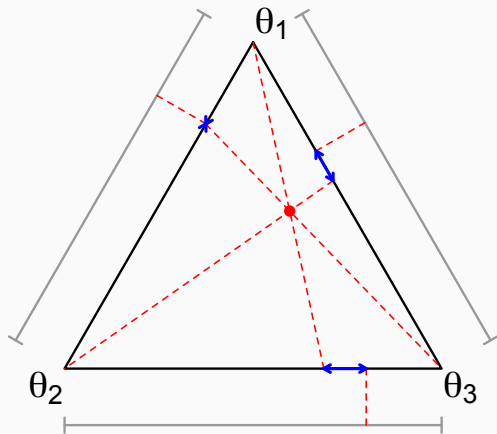
- $\hat{\theta}$: current estimate
- construct $P_{\mathcal{D}_{ij}}^{(b)}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$
- construct $P_{\theta_{ij}}^{(b)}$ from $\theta_{ij} = \{\theta_i, \theta_j\}$
- compare all the possible pairs



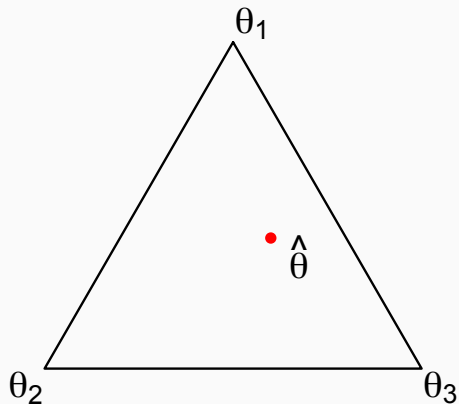
- initialize parameter



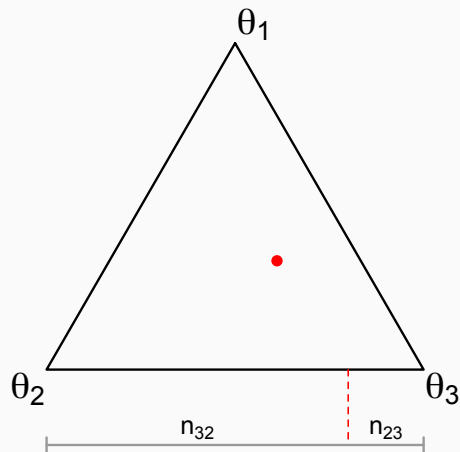
- initialize parameter
- update parameter to reduce total loss



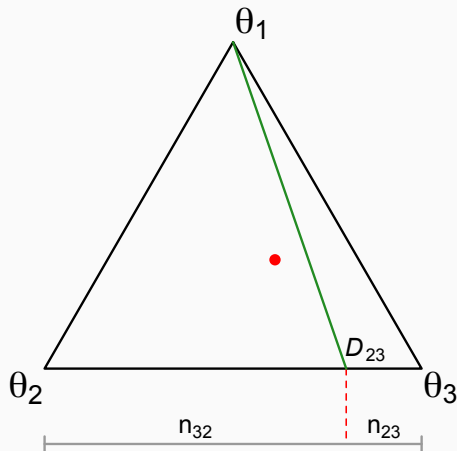
- initialize parameter
- update parameter to reduce total loss



- $\hat{\theta}$: current estimate



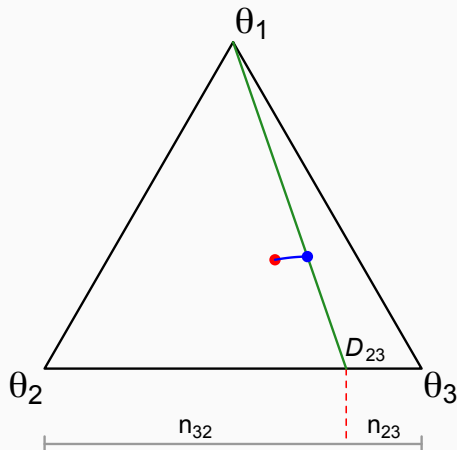
- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,



- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

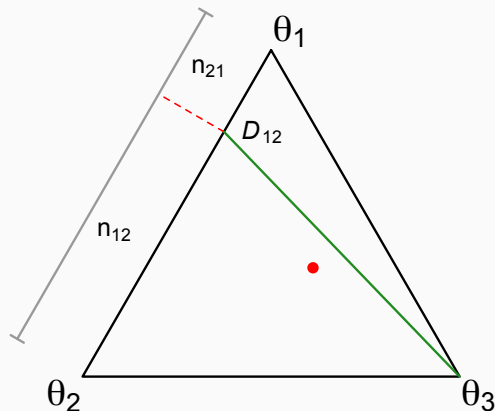


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij}
from $\hat{\theta}$



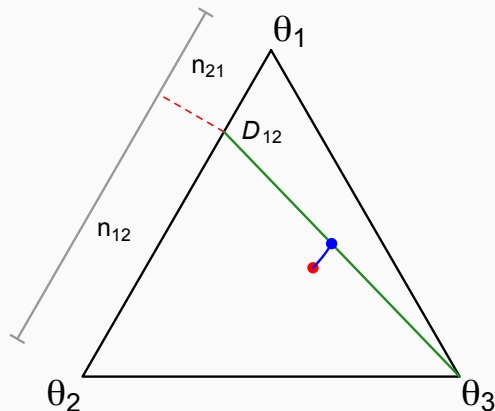
- $\hat{\theta}$: current estimate

- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$

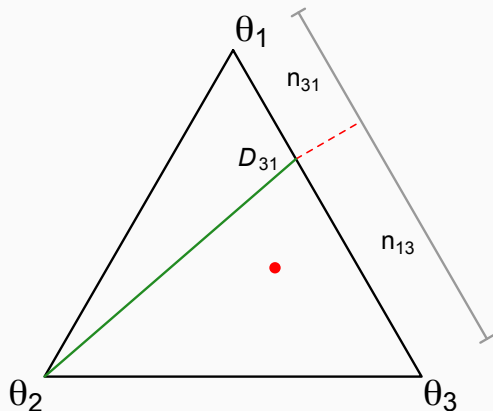


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$, consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$



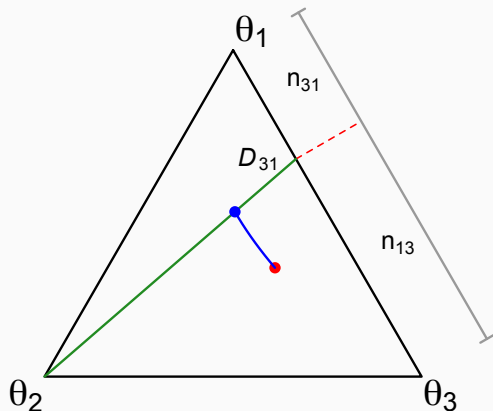
- $\hat{\theta}$: current estimate

- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$

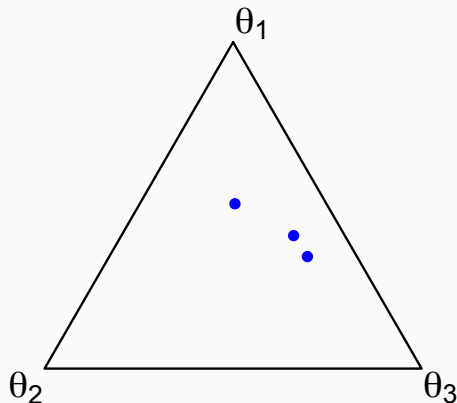


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij}
from $\hat{\theta}$

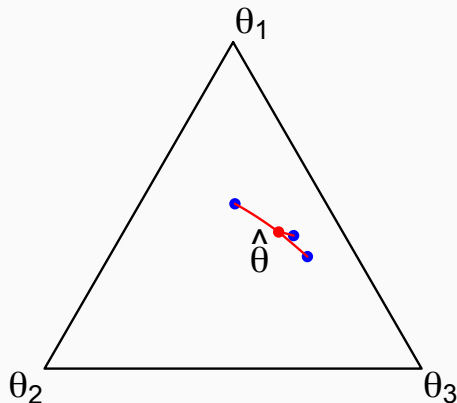


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij}
from $\hat{\theta}$

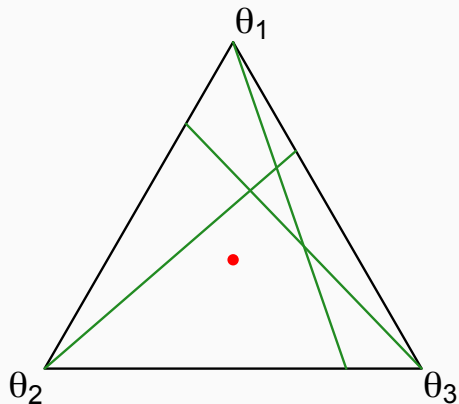


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,
consider a set of θ 's

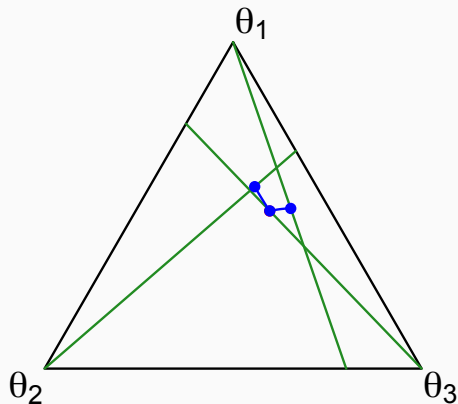
$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\},$$

which are consistent with pairwise comparison

- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$
- obtain $\hat{\theta}$ by integrating all $\tilde{\theta}_{ij}$'s

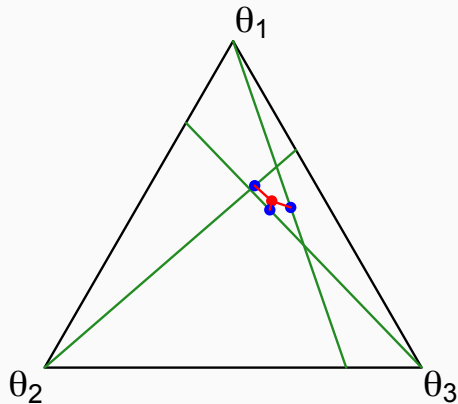


- initialize parameter



- initialize parameter
- e -projection:

$$\tilde{\theta}_{ij} = \arg \min_{\theta \in D_{ij}} D(P_{\theta}, P_{\hat{\theta}})$$



- initialize parameter
- e-projection:

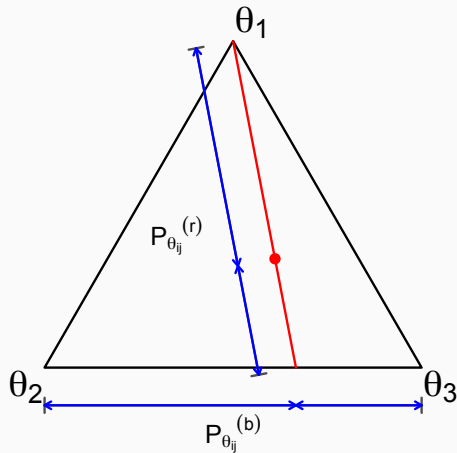
$$\tilde{\theta}_{ij} = \arg \min_{\theta \in D_{ij}} D(P_{\theta}, P_{\hat{\theta}})$$

- m -projection:

$$\hat{\theta} = \arg \min_{\theta} \sum_{i,j} w_{ij} D(P_{\tilde{\theta}_{ij}}, P_{\theta})$$

where $w_{ij} = (n_{ij} + n_{ji})$

DECOMPOSITION OF MULTINOMIAL DISTRIBUTION



$$P(\theta) = P_{\theta_{ij}}^{(b)} \times P_{\theta_{ij}}^{(r)}$$

- $P_{\theta_{ij}}^{(b)}$: binomial distribution on i and j
- $P_{\theta_{ij}}^{(r)}$: multinomial distribution on $\{i, j\}$ and the rest

- conventional method

$$\hat{\theta} = \arg \min_{\theta} \sum_{i < j} w_{ij} D(P_{\mathcal{D}_{ij}}^{(b)}, P_{\theta_{ij}}^{(b)})$$

- geometrical method

$$\hat{\theta} = \arg \min_{\theta} \sum_{i < j} w_{ij} D(P_{\mathcal{D}_{ij}}^{(b)}, P_{\theta_{ij}}^{(b)}) + \sum_{i < j} w'_{ij} D(P_{\mathcal{D}_{ij}}^{(r)}, P_{\theta_{ij}}^{(r)})$$

- this objective has a unique solution
- the second term works as a regularization

ILLUSTRATIVE EXAMPLE

Example from Hastie & Tibshirani (1998)

	1	2	3	4
1	-	0.56	0.51	0.60
2	0.44	-	0.96	0.44
3	0.49	0.04	-	0.59
4	0.40	0.56	0.41	-

Example from Hastie & Tibshirani (1998)

	1	2	3	4
1	-	0.56	0.51	0.60
2	0.44	-	0.96	0.44
3	0.49	0.04	-	0.59
4	0.40	0.56	0.41	-

- conventional estimates:

$$\{\hat{\theta}_i\} = \{0.29, 0.34, 0.16, 0.21\}$$

Example from Hastie & Tibshirani (1998)

	1	2	3	4
1	-	0.56	0.51	0.60
2	0.44	-	0.96	0.44
3	0.49	0.04	-	0.59
4	0.40	0.56	0.41	-

- conventional estimates:

$$\{\hat{\theta}_i\} = \{0.29, 0.34, 0.16, 0.21\}$$

- geometrical estimates:

$$\{\hat{\theta}_i\} = \{0.32, 0.29, 0.15, 0.23\}$$

- conventional estimates: $\{0.29, 0.34, 0.16, 0.21\}$
- geometrical estimates: $\{0.32, 0.29, 0.15, 0.23\}$

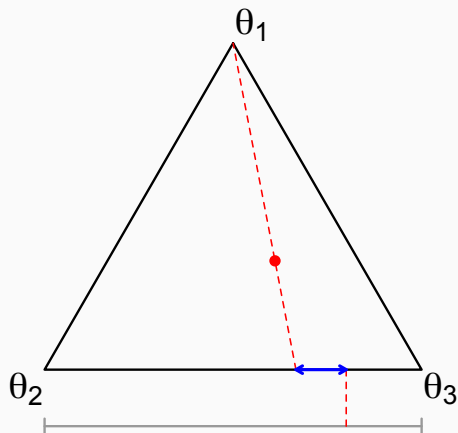
i	j	$P(i \succ j)$	$P(j \succ i)$	majority rule	conv.	geom.
1	2	0.56	0.44	$1 \succ 2$	\times	\checkmark
1	3	0.51	0.49	$1 \succ 3$	\checkmark	\checkmark
1	4	0.60	0.40	$1 \succ 4$	\checkmark	\checkmark
2	3	0.96	0.04	$2 \succ 3$	\checkmark	\checkmark
2	4	0.44	0.56	$2 \prec 4$	\times	\times
3	4	0.59	0.41	$3 \succ 4$	\times	\times

- generic form of objective

$$L(\theta) = \sum_{i < j} w_{ij} D(P_{\mathcal{D}_{ij}}, P_{\theta})$$

- weight w_{ij} reflects confidence of data \mathcal{D}_{ij}
- possible weights
 - data size of pairwise comparison
 - empirical influence of data
 - etc

proposal in Hastie & Tibshirani (1998)

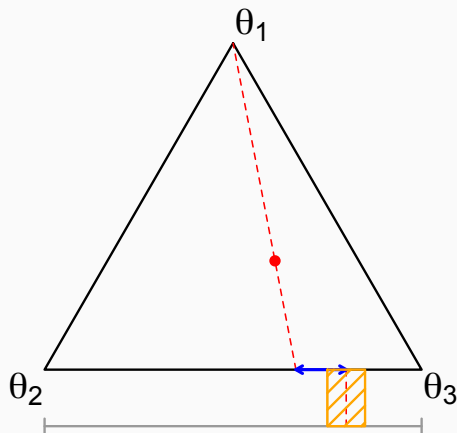


- binomial influence

$$w_{ij} \rightarrow \frac{w_{ij}}{\alpha(1-\alpha)}$$

$$\alpha = \frac{n_{ij}}{n_{ij} + n_{ji}}$$

proposal in Hastie & Tibshirani (1998)

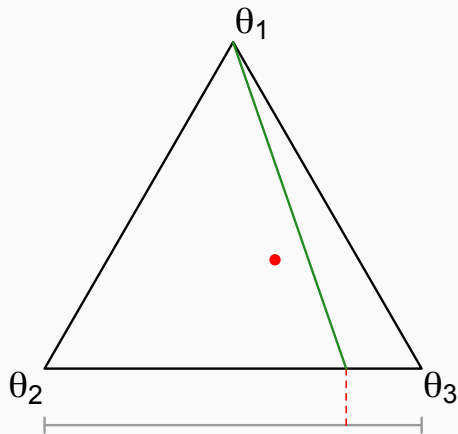


- binomial influence

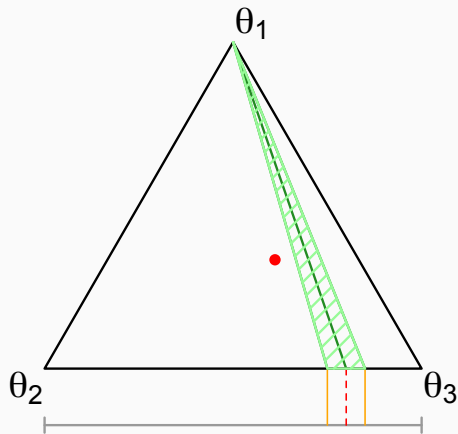
$$w_{ij} \rightarrow \frac{w_{ij}}{\alpha(1 - \alpha)}$$

$$\alpha = \frac{n_{ij}}{n_{ij} + n_{ji}}$$

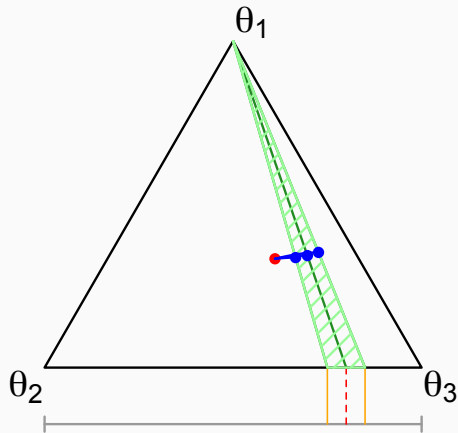
- weights are renormalized so as to equalize influences from variances of pairwise comparisons



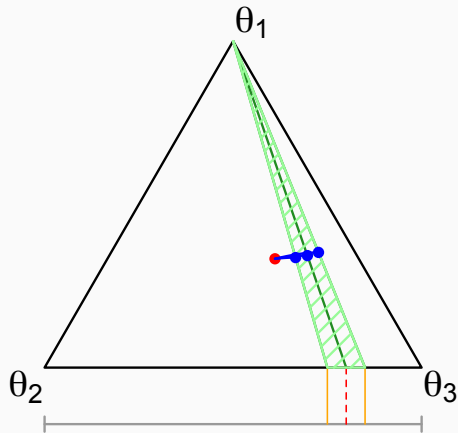
- influence around $\hat{\theta}$ should be considered



- influence around $\hat{\theta}$ should be considered
- fluctuation of data manifold



- influence around $\hat{\theta}$ should be considered
- fluctuation of data manifold
- fluctuation along e -geodesic is regarded as essential influence



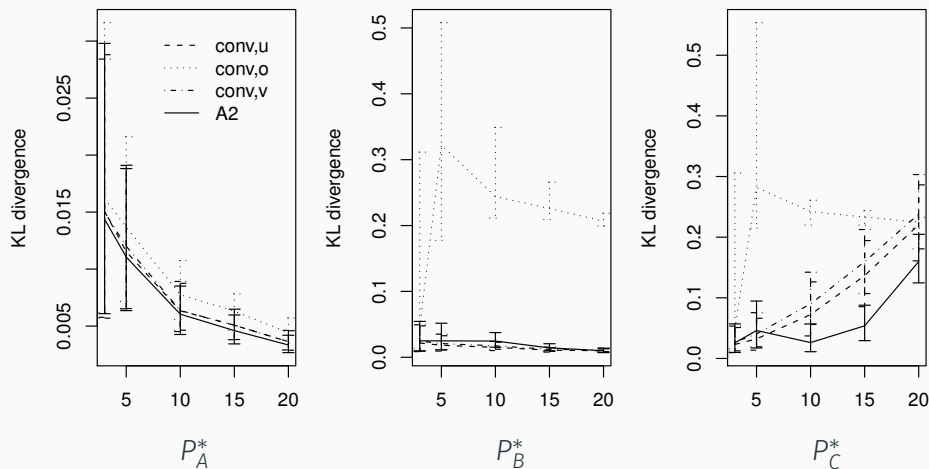
- influence around $\hat{\theta}$ should be considered
- fluctuation of data manifold
- fluctuation along e -geodesic is regarded as essential influence
- weights are determined so as to equalize those influences with iterative manner

Synthetic data in Hastie & Tibshirani (1998)

$$P_A^* = \left\{ \pi_i^* \mid \pi_1^* = \frac{1.5}{k}, \pi_j^* = \frac{1 - \pi_1^*}{k - 1} \ (j = 2, \dots, k) \right\}$$

$$P_B^* = \left\{ \pi_i^* \left| \begin{array}{l} \pi_1^* = \frac{2.85}{k}, \\ \pi_j^* = \frac{0.95 - \pi_1^*}{k/2 - 1} \quad \left(j = 2, \dots, \frac{k}{2} \right), \pi_j^* = \frac{0.05}{k/2} \quad \left(j = \frac{k}{2} + 1, \dots, k \right) \end{array} \right. \right\}$$

$$P_C^* = \left\{ \pi_i^* \left| \pi_1^* = 0.7125, \pi_2^* = 0.2375, \pi_j^* = \frac{0.05}{k-2} (j = 3, \dots, k) \right. \right\}$$



plots of the number of individuals vs. $D(P^*, P_{\hat{\theta}})$ for 500 trials.
 (solid:proposed, dashed:unit, dotted:# of data, dotdash:H&T)

Movie Rating Data

	Toy Story	Star Wars	Braveheart	The Saint	...
Anne	4	5		3	
Bob		5	4	2	
Cathy	5			3	
David	3	4	3	3	
⋮					

Movie Rating Data

	Toy Story	Star Wars	Braveheart	The Saint	...
Anne	4	5		3	
Bob		5	4	2	
Cathy	5			3	
David	3	4	3	3	
⋮					

characteristics of data

- each *user* gives a rate to each *item*
- some rates are not available
- rates are relative values, not absolute evaluation

Movie Rating Data

	Toy Story	Star Wars	Braveheart	The Saint	...
Anne	4	5		3	
Bob		5	4	2	
Cathy	5			3	
David	3	4	3	3	
⋮					

Problem

- estimate preference levels of items quantitatively
- predict preference levels of unrated items

marginalize with respect to hidden ordering (Hino, Fujimoto, and Murata 2010)

- observed ranking

$$\{i = j \succ \dots \succ k\}$$

- possible hidden ordering (unobserved)

$$\{i \succ j \succ \dots \succ k\} \text{ or } \{j \succ i \succ \dots \succ k\}$$

- marginalize with possible ordering

$$P(i = j \succ \dots \succ k) = P(i \succ j \succ \dots \succ k) + P(j \succ i \succ \dots \succ k)$$

notations:

- R_i^n : a rate of item i evaluated by user n
- $\mathcal{D}^n = \{R_1^n, R_2^n, \dots\}$: a set of rates given by user n
- $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \dots\}$: all data
- θ_i : preference parameter for item i
 - $\theta_i \geq 0$ (positivity)
 - $\sum \theta_i = 1$ (normalized)
- $\mathcal{S}(\mathcal{D}^n)$: a set of possible permutations for \mathcal{D}^n

- likelihood:

$$\begin{aligned}
 P(\mathcal{D}) &= \sum_n \sum_{\pi \in \mathcal{S}(\mathcal{D}^n)} P(\pi) P(\mathcal{D}^n | \pi) \\
 &= \sum_n \sum_{\pi \in \mathcal{S}(\mathcal{D}^n)} P(\pi) \prod_{i \in \pi} \frac{\theta_i}{\sum_{j \leq i \in \pi} \theta_j}
 \end{aligned}$$

where $P(\pi)$ is a prior of permutations

(marginalized with respect to all the possible ranking in equivalently rated items)

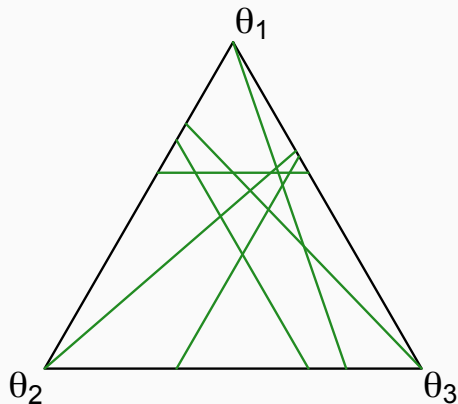
- the number of permutations increases with the number of items
exponentially

- decompose the objective into small optimization problems:

$$\text{minimize } \sum_r |\Lambda_r^n| \log \sum_{s \geq r} \Theta_s^n \quad \text{subject to } \sum_r \Theta_r^n = 1$$

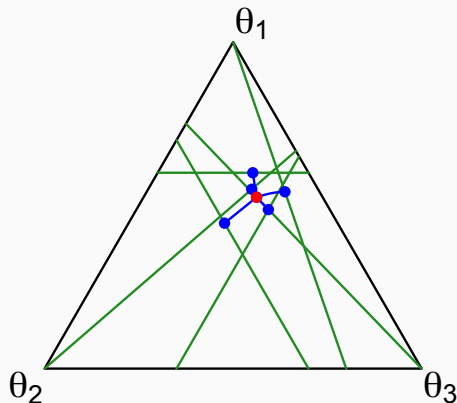
$$\text{maximize } \sum_i \log \theta_i \quad \text{subject to } \sum_i \theta_i = 1$$

- algorithm
 - find solutions of the minimization problems
 - find a parameter of the maximization problem which is as consistent with those solutions as possible



- solutions of minimization problems

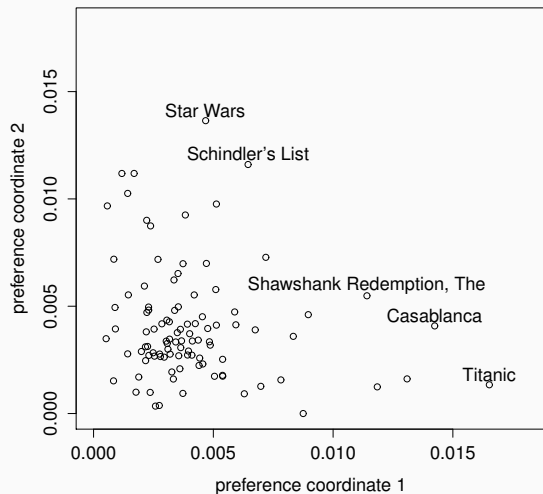
$$\mathcal{D}^n = \{\theta \mid \sum_{i \in \Lambda_r^n} \theta_i = \text{const.}\}$$



- solutions of minimization problems

$$\mathcal{D}^n = \{\theta \mid \sum_{i \in \Lambda_r^n} \theta_i = \text{const.}\}$$

- find a estimate with geometrical BT method



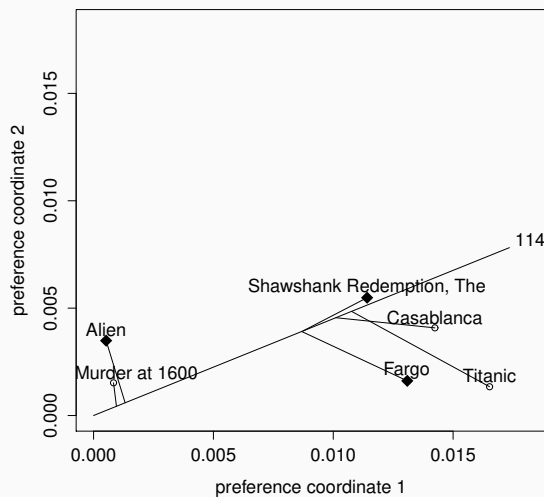
- preference parameters are modeled as

$$\theta_{iu} = v_i \cdot w_u$$

$v_i \in \mathbb{R}^d$: item i ,

$w_u \in \mathbb{R}^d$: user u

- \circ movies
- typical two axes are used



- ○ rated by user~114
- ◇ not rated





CONCLUSION

we presented the following

- a geometrical reformulation of the estimation procedure for the Bradley-Terry model
- a robust weight adaptation method
- an approximate estimation for grouped ranking data

in addition, possible application would be

- utilizing U -divergence based on m -flat nature of data manifolds

-  Bradley, Ralph Allan and Milton E. Terry (Dec. 1952). “Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons.” In: *Biometrika* 39.3/4, pp. 324–345. DOI: [10.2307/2334029](https://doi.org/10.2307/2334029). JSTOR: [2334029](https://www.jstor.org/stable/2334029).
-  Fujimoto, Yu, Hideitsu Hino, and Noboru Murata (June 2011). “An Estimation of Generalized Bradley-Terry Models Based on the em Algorithm.” In: *Neural Computation* 23.6, pp. 1623–1659. DOI: [10.1162/NECO_a_00129](https://doi.org/10.1162/NECO_a_00129).
-  Hastie, Trevor and Robert Tibshirani (Apr. 1998). “Classification by Pairwise Coupling.” In: *The Annals of Statistics* 26.2, pp. 451–471. DOI: [10.1214/aos/1028144844](https://doi.org/10.1214/aos/1028144844). JSTOR: [120036](https://www.jstor.org/stable/120036).
-  Hino, Hideitsu, Yu Fujimoto, and Noboru Murata (Sept. 2010). “A Grouped Ranking Model for Item Preference Parameter.” In: *Neural Computation* 22.9, pp. 2417–2451. DOI: [10.1162/NECO_a_00008](https://doi.org/10.1162/NECO_a_00008).