A geometrical extension of the Bradley-Terry model

Information Geometry of ranking problem

Noboru Murata

September 19, 2021

Waseda University

Introduction

example: win-loss records of MLB (American East)

number of wins

	Y	R	RS	BJ	O
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)

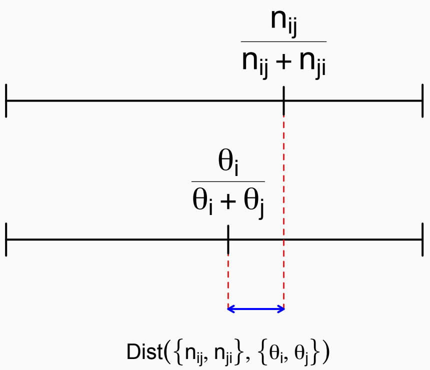
compare binomial distributions:

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

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

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

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



- discrepancy (KL divergence):

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

conventional algorithm: (Hastie and Tibshirani 1998)

- objectives: likelihood of binomial distribution

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

- 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

basic ideas: (Fujimoto, Hino, and Murata 2011)

- BT model parameter can be identified with a multinomial distribution

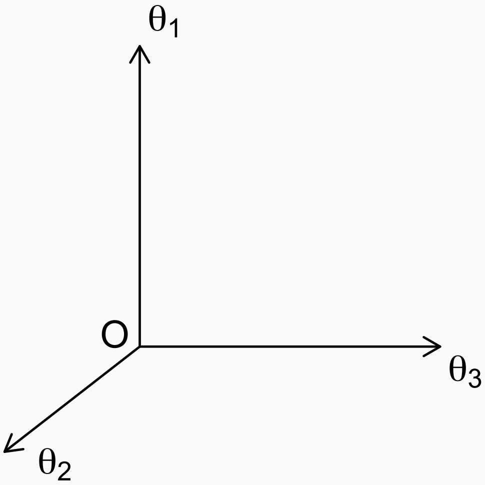
  a point on the probability simplex
- pairwise comparison data can be regarded as incomplete data from multinomial distributions

  an *m*-flat manifold in the probability simplex

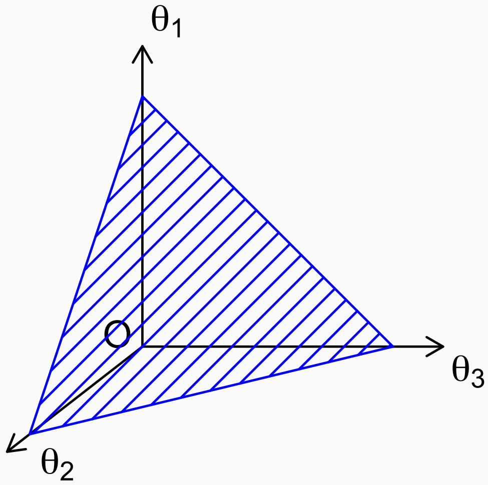
_em_ algorithm (Amari, 1995)

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

example: $k = 3$

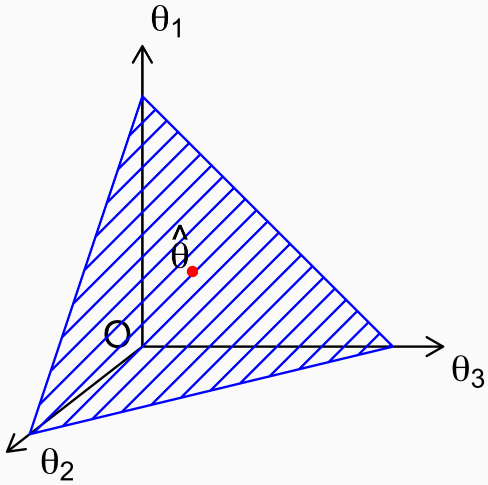


example: $k = 3$

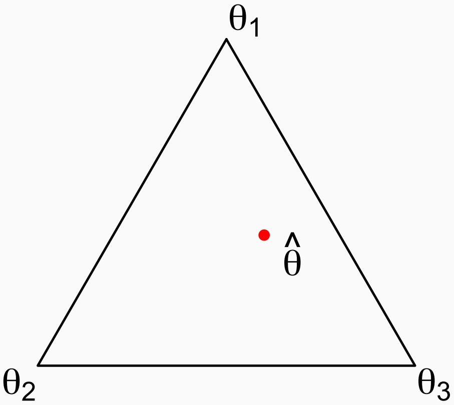


- $\theta_i \geq 0$ (positivity)
- $\sum \theta_i = 1$ (normalized)

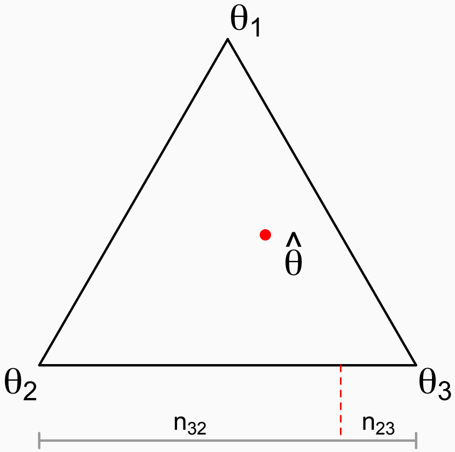
example: $k = 3$



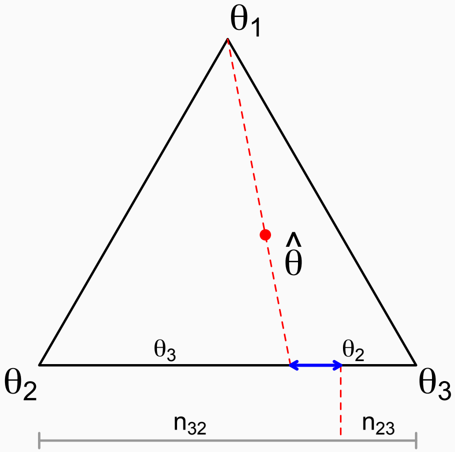
- $\theta_i \geq 0$ (positivity)
- $\sum \theta_i = 1$ (normalized)
- estimate $\hat{\theta}$:
  a point in the simplex



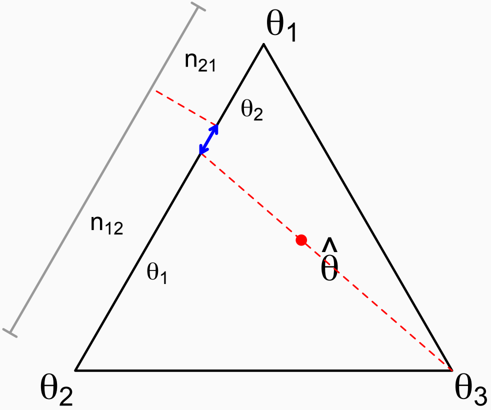
- $\hat{\theta}$: current estimate



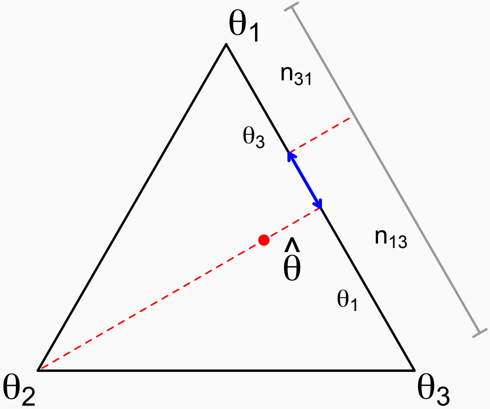
- $\hat{\theta}$: current estimate
- construct $P^{(b)}_{\mathcal{D}_{ij}}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$



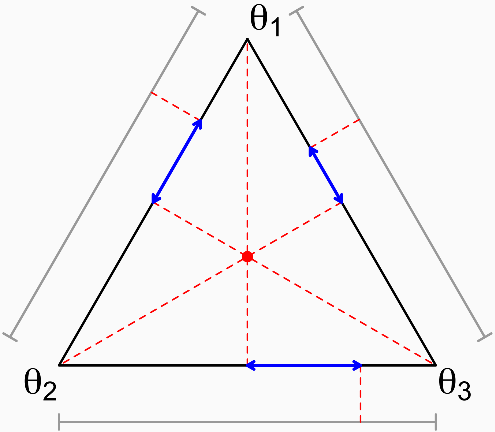
- $\hat{\theta}$: current estimate
- construct $P^{(b)}_{\mathcal{D}_{ij}}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$
- construct $P^{(b)}_{\theta_{ij}}$ from $\theta_{ij} = \{\theta_i, \theta_j\}$



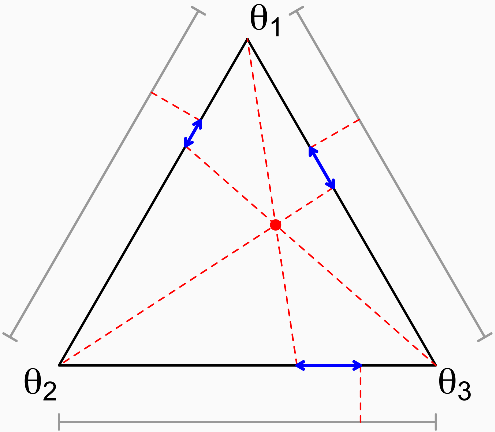
- $\hat{\theta}$: current estimate
- construct $P^{(b)}_{\mathcal{D}_{ij}}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$
- construct $P^{(b)}_{\theta_{ij}}$ from $\theta_{ij} = \{\theta_i, \theta_j\}$
- compare all the possible pairs



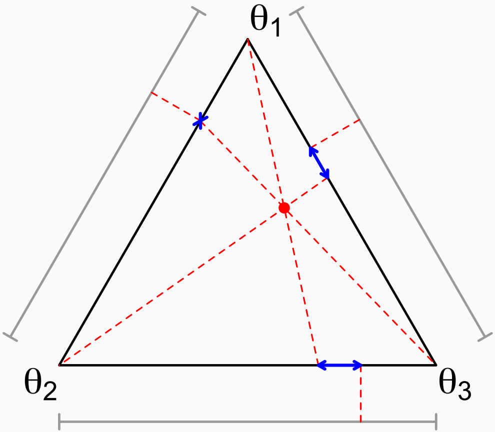
- $\hat{\theta}$: current estimate
- construct $P^{(b)}_{\mathcal{D}_{ij}}$ from $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$
- construct $P^{(b)}_{\theta_{ij}}$ from $\theta_{ij} = \{\theta_i, \theta_j\}$
- compare all the possible pairs



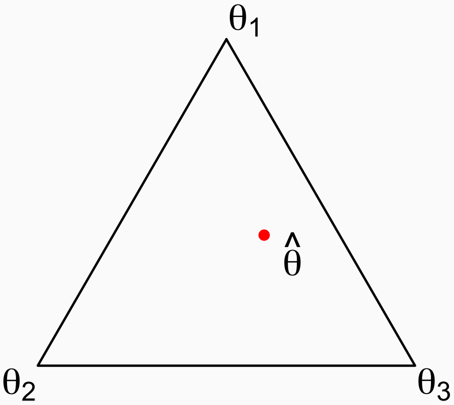
- initialize parameter



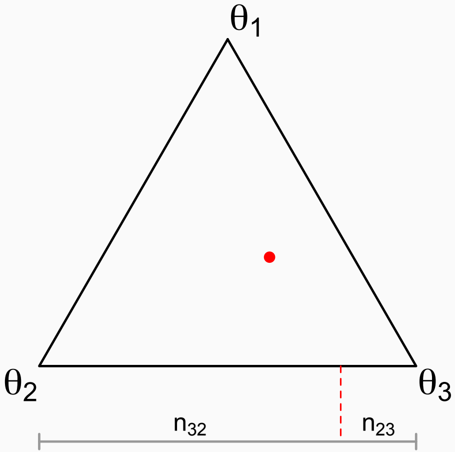
- initialize parameter
- update parameter to reduce the total loss



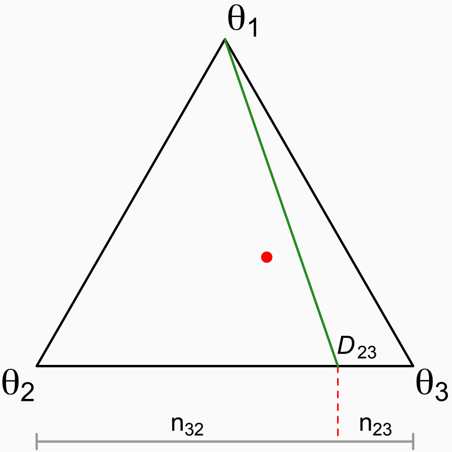
- initialize parameter
- update parameter to reduce the total loss



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

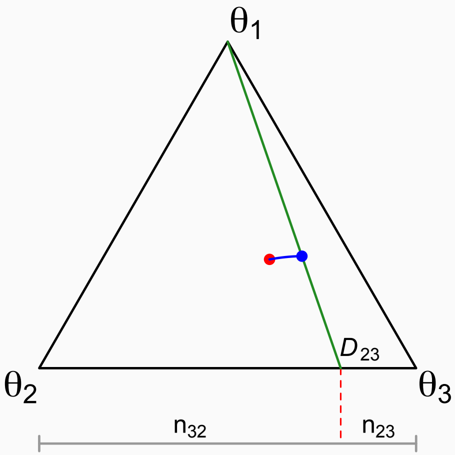


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,



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

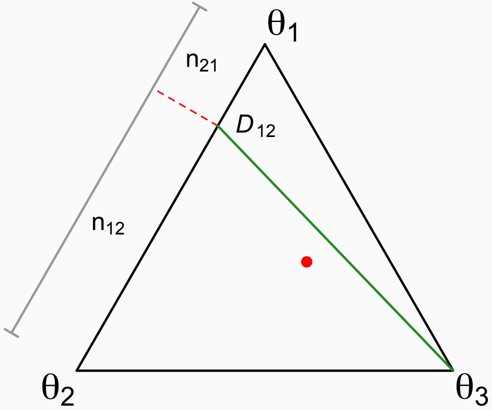
$$D_{ij} = \{\theta | \theta_i : \theta_j = n_{ij} : n_{ji}\}$$



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

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

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

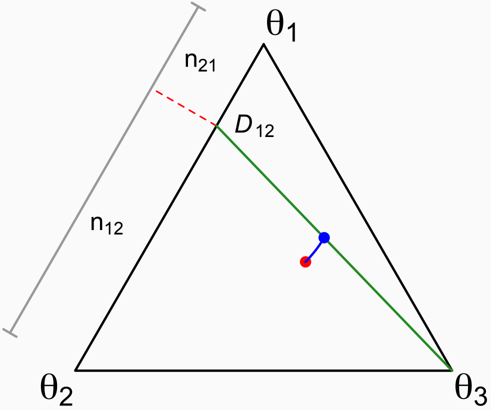


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,

  consider a set of consistent θ

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

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

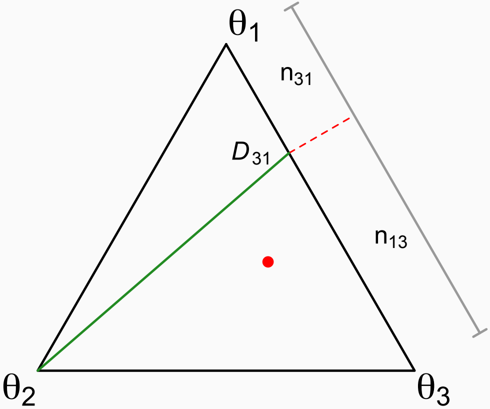


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,

  consider a set of consistent θ

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

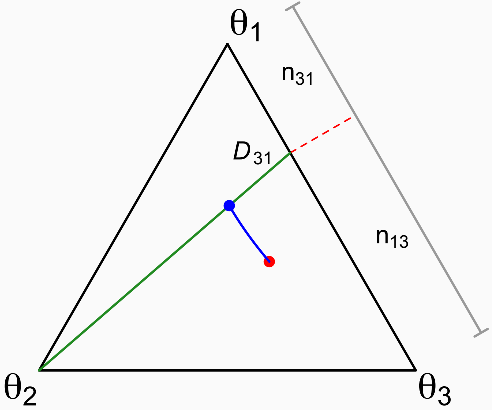
- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$



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

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

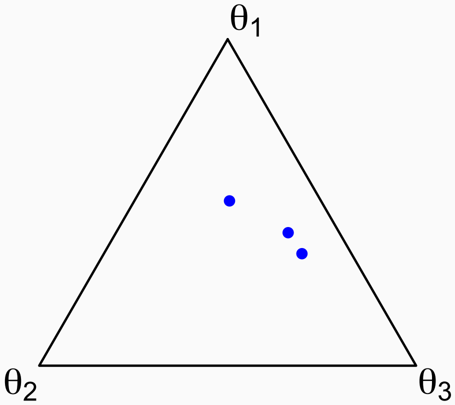
- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$



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

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

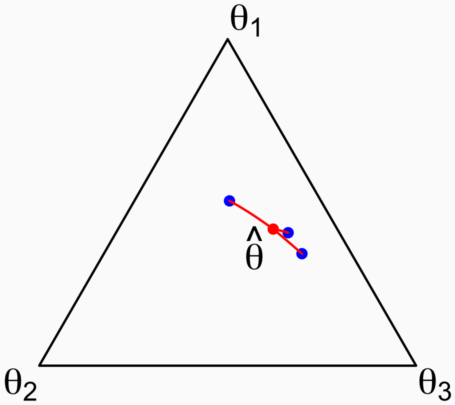
- choose the closest point $\tilde{\theta}_{ij}$ in D_{ij} from $\hat{\theta}$



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

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

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

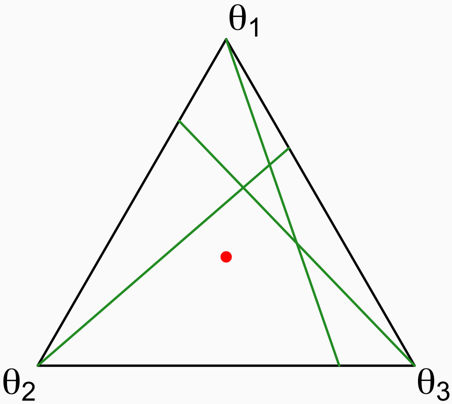


- $\hat{\theta}$: current estimate
- for $\mathcal{D}_{ij} = \{n_{ij}, n_{ji}\}$,

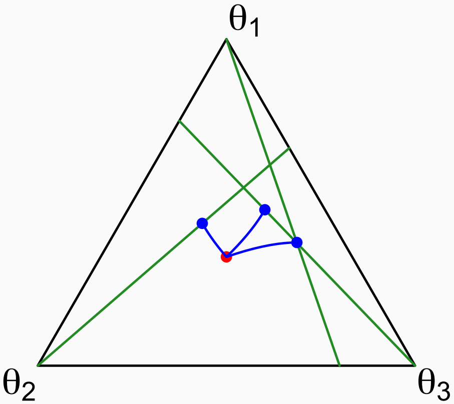
  consider a set of consistent θ

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

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

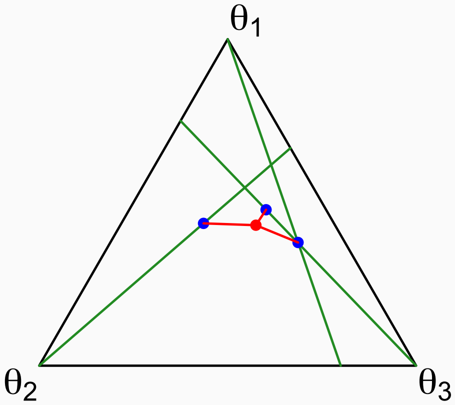


- initialize parameter



- initialize parameter
- *e*-projection:

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

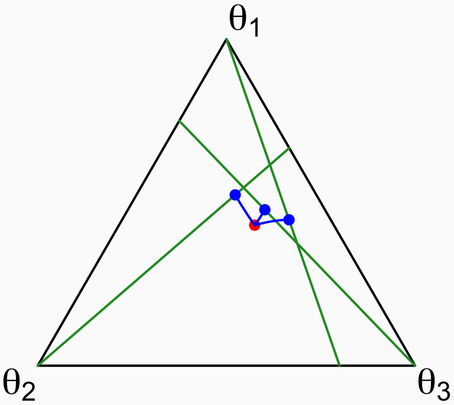


- initialize parameter

- *m*-projection:

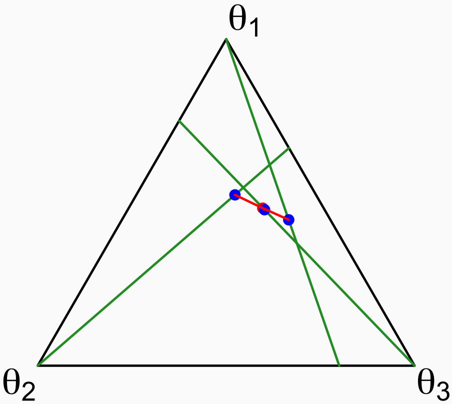
$$\hat{\theta} = \arg\min_{\theta} \sum_{i,j} w_{ij} D(P_{\tilde{\theta}_{ij}}, P_{\theta})$$

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



- initialize parameter
- *e*-projection:

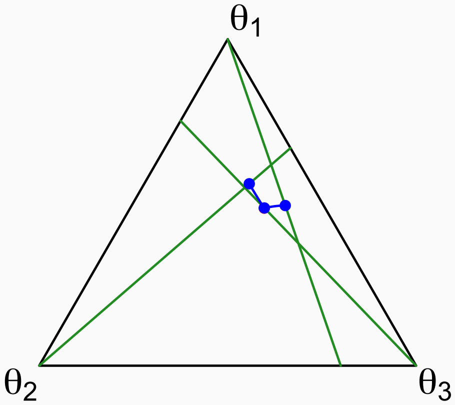
$$\tilde{\theta}_{ij} = \arg\min_{\theta \in D_{ij}} D(P_\theta, P_{\hat{\theta}})$$



- initialize parameter
- *m*-projection:

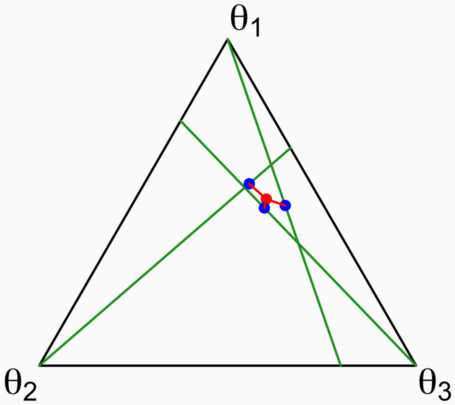
$$\hat{\theta} = \arg\min_{\theta} \sum_{i,j} w_{ij} D(P_{\tilde{\theta}_{ij}}, P_{\theta})$$

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



- initialize parameter
- *e*-projection:

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

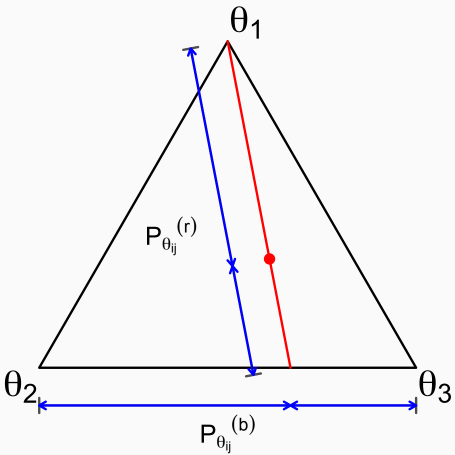


- initialize parameter

- m-projection:

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

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



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

- $P_{\theta_{ij}}^{(b)}$: binomial distribution on i and j
- $P_{\theta_{ij2}}^{(r)}$: multinomial distribution on $\{i, j\}$ and the rest

conventional method:

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

geometrical method:

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

- 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	-

- conventional estimate:

$$\{\theta_i\} = \{0.29, 0.34, 0.16, 0.21\}$$

- geometrical estimate:

$$\{\theta_i\} = \{0.32, 0.29, 0.15, 0.23\}$$

- conv. est.: $\{0.29, 0.34, 0.16, 0.21\}$
- geom. est.: $\{0.32, 0.29, 0.15, 0.23\}$

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

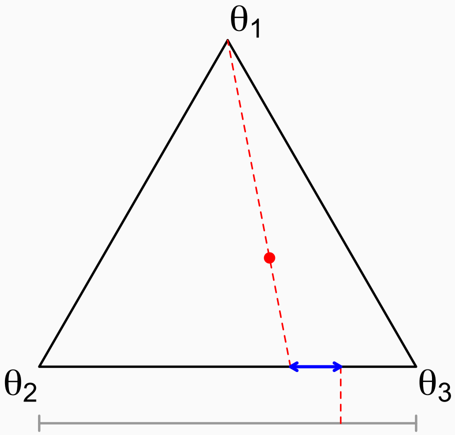
weights in objective:

- 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
    - ...

a proposal in Hastie & Tibshirani (1998)

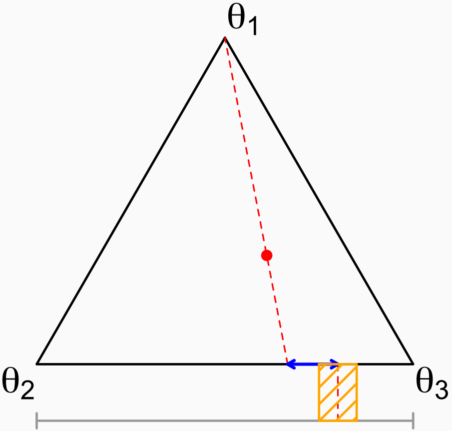


- binomial influence

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

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

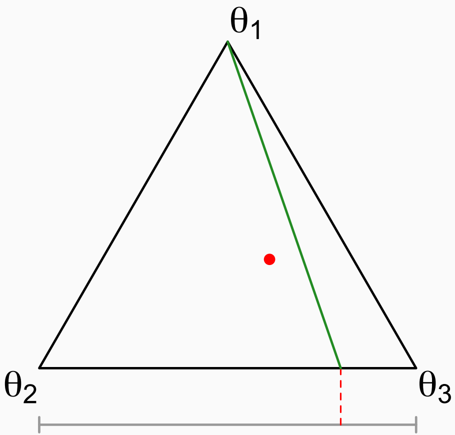
a proposal in Hastie & Tibshirani (1998)



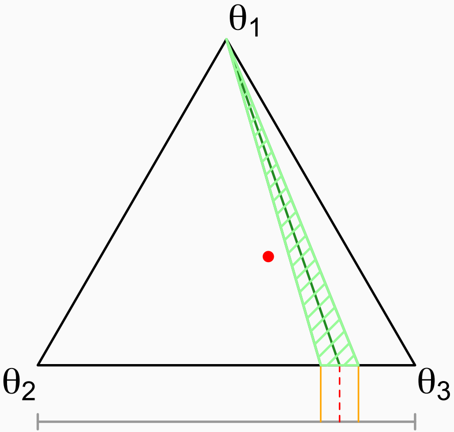
- binomial influence

$$w_{ij} \to \frac{w_{ij}}{\alpha(1-\alpha)}$$
$$\alpha = \frac{n_{ij}}{n_{ij}+n_{ji}}$$

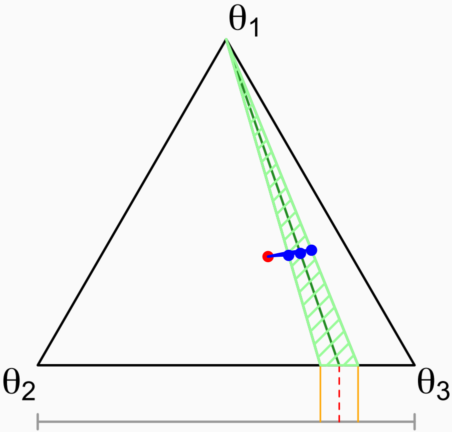
- weights are renormalized so as to equalize influences from variances of pairwise comparisons



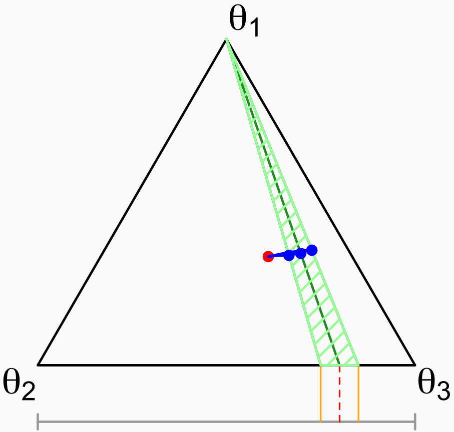
- influence around $\hat{\theta}$ should be considered



- 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



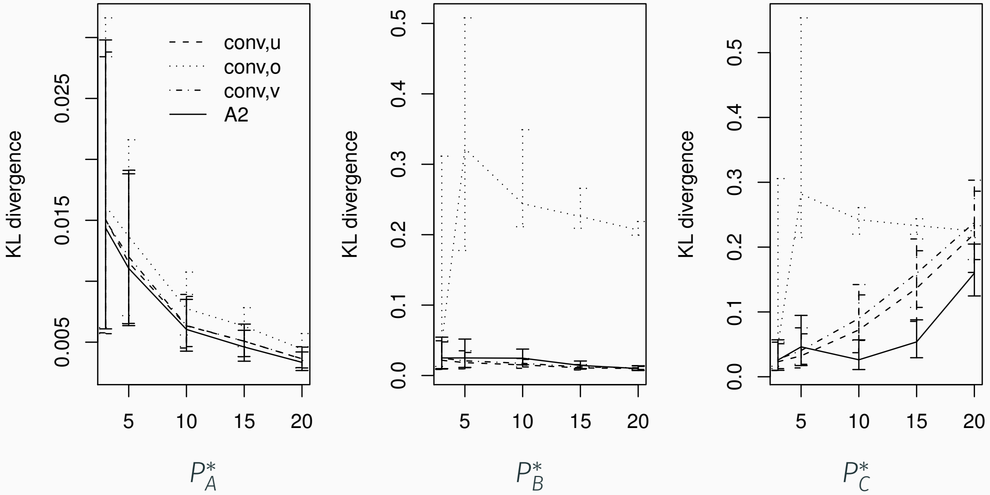
- influence around $\hat{\theta}$ should be considered
- fluctuation of data manifold
- fluctuation along *e*-geodesic is regarded as essential influence
- weights are determined so as to equalize those influences with iterative manner

synthetic data in Hastie & Tibshirani (1998)

$$P_A^* = \left\{ \pi_i^* \;\middle|\; \pi_1^* = \frac{1.5}{k}, \pi_j^* = \frac{1-\pi_1^*}{k-1} \; (j = 2, \ldots, k) \right\}$$

$$P_B^* = \left\{ \pi_i^* \;\middle|\; \pi_1^* = \frac{2.85}{k}, \pi_j^* = \frac{0.95-\pi_1^*}{k/2-1} \; \left(j = 2, \ldots, \frac{k}{2}\right) \right.$$
$$\left. \pi_j^* = \frac{0.05}{k/2} \; \left(j = \frac{k}{2}+1, \ldots, k\right) \right\}$$

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



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)

example: 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

problem:

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

marginalize with respect to hidden ordering:

- observed ranking

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

- possible hidden ordering (unobserved)

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

- marginalize with possible ordering

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

(Hino, Fujimoto, and Murata 2010)

notations:

- R_i^n: a rate of item i evaluated by user n
- $\mathcal{D}^n = \{R_1^n, R_2^n, \ldots\}$: a set of rates given by user n
- $\mathcal{D} = \{\mathcal{D}^1, \mathcal{D}^2, \ldots\}$: 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:

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

  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 basic idea:

- bound below with tractable calculation
- exclude marginalization of permutation

lower bound of likelihood:

- $\Lambda_r^n = \{i | R_i^n = r\}$: an index set of items with rate r by user n
- $\Theta_r^n = \sum_{i \in \Lambda_r^n} \theta_i$: group preference parameter (total preference of the equivalently rated items)

$$\underline{P}(\mathcal{D}) = \sum_n \prod_r \prod_{i \in \Lambda_r^n} \frac{\theta_i}{\sum_{s \leq r} \Theta_s^n}$$

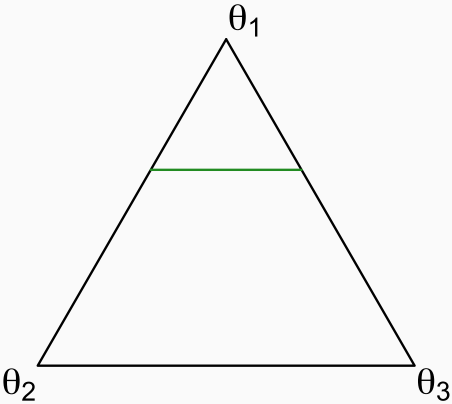
maximizing lower bound is still computationally tedious decompose the objective into small optimization problems:

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

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

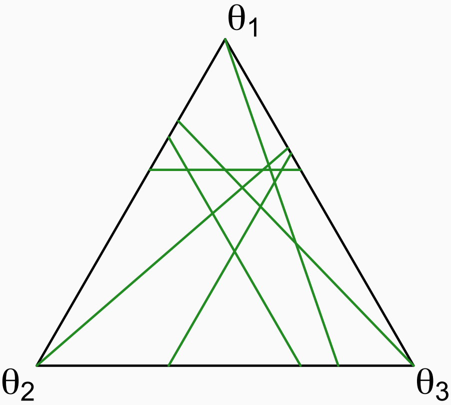
algorithm:

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



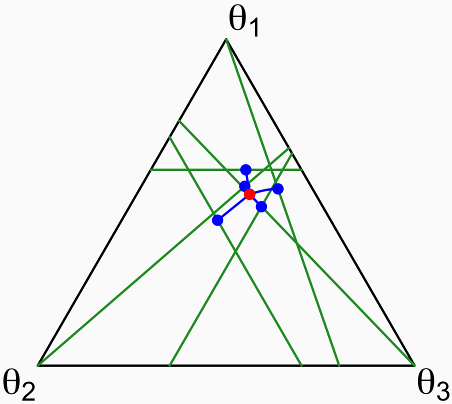
- solutions of minimization problems

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



- solutions of minimization problems

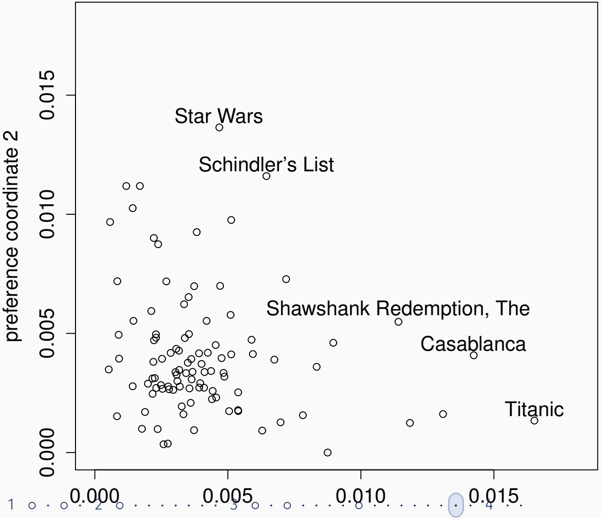
$$\mathcal{D}^n = \{\theta | \sum_{i \in \Lambda_r^n} \theta_i = const.\}$$



- solutions of minimization problems

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

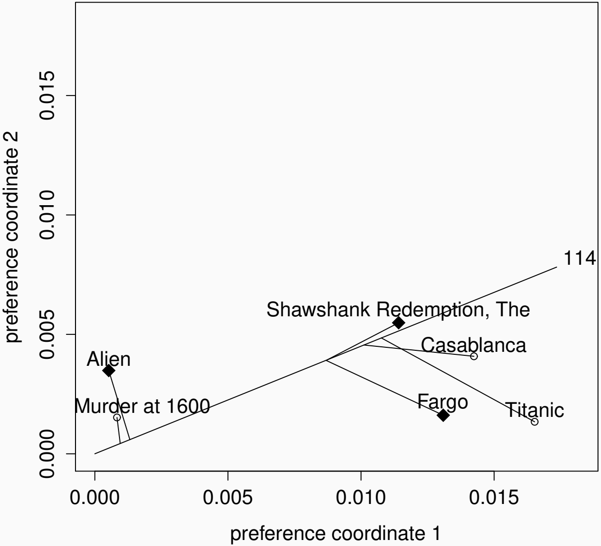
- find a estimate with geometrical BT method application of mixture model:



item mapping:

typical two axes are used

- movies



item-user mapping:

- ○ rated by user 114
- ◇ not rated

Concluding Remarks

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. JSTOR: 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.

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. JSTOR: 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.