Graph Helmholtzian and Rank Learning

Lek-Heng Lim

NIPS Workshop on Algebraic Methods in Machine Learning

December 12, 2008

(Joint work with Xiaoye Jiang, Yuan Yao, and Yinyu Ye)

Modern ranking data

- Multicriteria decision systems
    - Recommendation system (user-product, e.g. Amazon)
    - Interest ranking in social networks (person-interest, e.g. LinkedIn)
    - Popularity contest (voter-candidate, e.g. YouTube)
- Peer review systems
    - publication citation systems (paper-paper, e.g. CiteSeer)
    - webpage ranking (web-web, e.g. Google)
    - reputation system (customer-customer, e.g. eBay)
- Alternatives: websites, scholarly articles, sellers, movies
- Voters:
    - other websites, other scholarly articles, buyers, viewers
    - groups of websites (topics), scholarly articles (authorship), buyers (buying pattern), viewers (movie taste)
    - different criteria used to judge the alternatives

Old problems with ranking

- Condorcet's paradox: intransitivity can happen in group decision making, i.e. the majority prefers a to b and b to c, but may yet prefer c to a.
    - [Condorcet, 1785]
- Impossibility theorems in social choice: any societal preference aggregation that is sophisticated enough must exhibit intransitivitiy.
    - [Arrow, 1950]
    - [Sen, 1970]
- Empirical studies in psychology:
    - lack of majority consensus common in group decision making,
    - even an individual can exhibit such seemingly irrational behaviour (multiple criteria used to make preference judgement).

New problems with ranking

Modern ranking data often

- incomplete: typically about 1%,
- imbalanced: power-law, heavy-tail distributed votes,
- cardinal: given in terms of scores or stochastic choices.

Implicitly or explicitly, ranking data may be regarded to be living on a **pairwise comparison graph** $G = (V, E)$, where

- V: set of alternatives (products, interests, etc) to be ranked,
- E: pairs of alternatives to be compared.

Properties

- incomplete: sparsity in E,
- imbalanced: degree distribution of V,
- cardinal: real-valued functions on V.

Example: Netflix customer-product rating

Example (Netflix Customer-Product Rating)

- 480189-by-17770 customer-product rating matrix A.
- **incomplete**: 98.82% of values missing.
- **imbalanced**: number of raters on each movie varies enormously.

However,

- pairwise comparison graph $G = (V, E)$ is very **dense**!
- only 0.22% edges are missed, **almost a complete graph**
- rank aggregation may be carried out without estimating missing values

Caveat: we are not trying to solve the Netflix prize problem

Objective

Ranking data in form of voter-alternative ratings $A = [a_{\alpha i}]$, highly incomplete and imbalanced. Want the following.

- A global ranking of the alternatives if one exists.
- A certificate of reliability to quantify the validity of the global ranking.
- If there are no meaningful global ranking, analyze cyclic inconsistencies. E.g. Are the inconsistencies local or global or neither?
- Allow for globally cyclic rankings.

Local inconsistencies

If there are only local inconsistencies, then

- Condorcet paradox happens to items ranked closed together but not to items ranked far apart, i.e. ordering of 4th, 5th, 6th ranked items cannot be trusted but ordering of 4th, 60th, 100th ranked items can;
- may rank groups of alternatives: e.g. among gourmets, no consensus for hamburgers, hot dogs, pizzas, and no consensus for caviar, foie gras, truffles, but clear preference for latter group.

Basic model for rank learning

Optimize over model class $\mathcal{M}$

$$\min_{X \in \mathcal{M}} \sum\nolimits_{\alpha,i,j} w_{ij}^{\alpha} (X_{ij} - Y_{ij}^{\alpha})^2.$$

Y_{ij}^{α} quantifies degree of preference of alternative i over alternative j held by voter α. Y^{α} skew-symmetric matrix.

$$w_{ij}^{\alpha} = w(\alpha, i, j) = \begin{cases} 1 & \text{if } \alpha \text{ made comparison for } \{i, j\}, \\ 0 & \text{otherwise.} \end{cases}$$

- Kemeny optimization:

$$\mathcal{M}_K := \{X \in \mathbb{R}^{n \times n} \mid X_{ij} = \operatorname{sign}(s_j - s_i), s : V \to \mathbb{R}\},$$

  where sign : $\mathbb{R} \to \{\pm 1\}$. NP-hard to compute.

- Relaxed version:

$$\mathcal{M}_G = \{X \in \mathbb{R}^{n \times n} \mid X_{ij} = s_j - s_i, s : V \to \mathbb{R}\}.$$

  Least squares regression over skew-symmetric matrices of rank 2.

Rank aggregation

- Previous problem may be reformulated

$$\min_{X \in \mathcal{M}_G} \|X - \bar{Y}\|_{2,w}^2 = \min_{X \in \mathcal{M}_G} \left[\sum\nolimits_{\{i,j\} \in E} w_{ij} (X_{ij} - \bar{Y}_{ij})^2 \right]$$

  where

$$w_{ij} := \sum\nolimits_\alpha w_{ij}^\alpha \quad \text{and} \quad \bar{Y}_{ij} := \frac{\sum_\alpha w_{ij}^\alpha Y_{ij}^\alpha}{\sum_\alpha w_{ij}^\alpha}.$$

- Why not just aggregate over scores directly? Mean score is a **first order** statistics and is inadequate because
  - most voters would rate just a very small portion of the alternatives,
  - different alternatives may have different voters, mean scores affected by individual rating scales.
- Use higher order statistics.

Pairwise rank aggregation

Given voter-alternative rating matrix $A = [a_{\alpha i}]$ (highly incomplete).

- **Linear Model**: average score difference between product i and j over all voters who have rated both of them,

$$\bar{Y}_{ij} = \frac{\sum_{\alpha}(a_{\alpha j} - a_{\alpha i})}{\#\{\alpha \mid a_{\alpha i}, a_{\alpha j} \text{ exist}\}}.$$

  Invariant up to translation.

- **Log-linear Model**: when all the scores are positive, the logarithmic average score ratio,

$$\bar{Y}_{ij} = \frac{\sum_{\alpha}(\log a_{\alpha j} - \log a_{\alpha i})}{\#\{\alpha \mid a_{\alpha i}, a_{\alpha j} \text{ exist}\}}.$$

  Invariant up to a multiplicative constant.

More second order statistics

- **Linear Probability Model**: the probability difference j is preferred to i than the other way round,

$$\bar{Y}_{ij} = \Pr\{\alpha \mid a_{\alpha j} > a_{\alpha i}\} - \Pr\{\alpha \mid a_{\alpha j} < a_{\alpha i}\}.$$

  Invariant up to monotone transformation.

- **Bradley-Terry Model**: logarithmic odd ratio (logit)

$$\bar{Y}_{ij} = \log \frac{\Pr\{\alpha \mid a_{\alpha j} \geq a_{\alpha i}\}}{\Pr\{\alpha \mid a_{\alpha j} \leq a_{\alpha i}\}}.$$

  Invariant up to monotone transformation.

Functions on graph

$G = (V, E)$ undirected graph. V vertices, $E \in \binom{V}{2}$ edges, $T \in \binom{V}{3}$ triangles/3-cliques. $\{i, j, k\} \in T$ iff $\{i, j\}, \{j, k\}, \{k, i\} \in E$.

- **Function on vertices:** $s : V \to \mathbb{R}$
- **Edge flows:** $X : V \times V \to \mathbb{R}$, $X(i, j) = 0$ if $\{i, j\} \notin E$,

$$X(i, j) = -X(j, i) \quad \text{for all } i, j.$$

- **Triangular flows:** $\Phi : V \times V \times V \to \mathbb{R}$, $\Phi(i, j, k) = 0$ if $\{i, j, k\} \notin T$,

$$\begin{aligned}\Phi(i, j, k) &= \Phi(j, k, i) = \Phi(k, i, j) \\ &= -\Phi(j, i, k) = -\Phi(i, k, j) = -\Phi(k, j, i) \quad \text{for all } i, j, k.\end{aligned}$$

- Physics: s, X, Φ potential, alternating vector/tensor field.
- Topology: s, X, Φ 0-, 1-, 2-cochain.
- Ranking: s scores/utility, X pairwise rankings, Φ triplewise rankings

Hilbert space of forms

- Let $|V| = n$. Then s is a vector in $\mathbb{R}^n$, X is a skew-symmetric matrix in $\mathbb{R}^{n\times n}$, Φ is an alternating 3-tensor in $\mathbb{R}^{n\times n\times n}$.
- Inner products

$$\langle s, t\rangle = \sum\nolimits_i w_i s_i t_i, \qquad \langle X, Y\rangle = \sum\nolimits_{i,j} w_{ij} X_{ij} Y_{ij},$$
$$\langle \Phi, \Psi\rangle = \sum\nolimits_{i,j,k} w_{ijk} \Phi_{ijk} \Psi_{ijk},$$

  - $w = [w_i] \in \mathbb{R}^n_+$, $w_i > 0$ all i;
  - $W = [w_{ij}] \in \mathbb{R}^{n\times n}_+$ symmetric, $w_{ij} = 0$ iff $\{i, j\} \notin E$,
  - $\mathcal{W} = [w_{ijk}] \in \mathbb{R}^{n\times n\times n}_+$ symmetric, $w_{ijk} = 0$ iff $\{i, j, k\} \notin T$.
- For simplicity, assume $w_i = 1$, $w_{ijk} = \mathbf{1}_T(\{i, j, k\})$. Write
  - $L^2(V) = L^2_w(V)$,
  - $L^2(E) = L^2_W(V \wedge V)$,
  - $L^2(T) = L^2_{\mathcal{W}}(V \wedge V \wedge V)$.

Operators

- **Gradient:** $\operatorname{grad} : L^2(V) \to L^2(E)$,
$$(\operatorname{grad} s)(i,j) = s_j - s_i.$$
- **Curl:** $\operatorname{curl} : L^2(E) \to L^2(T)$,
$$(\operatorname{curl} X)(i,j,k) = X_{ij} + X_{jk} + X_{ki}.$$
- **Divergence:** $\operatorname{div} : L^2(E) \to L^2(V)$,
$$(\operatorname{div} X)(i) = \sum\nolimits_j w_{ij} X_{ij}.$$
- **Graph Laplacian:** $\Delta_0 : L^2(V) \to L^2(V)$,
$$\Delta_0 = \operatorname{div} \circ \operatorname{grad}.$$
- **Graph Helmholtzian:** $\Delta_1 : L^2(E) \to L^2(E)$,
$$\Delta_1 = \operatorname{curl}^* \circ \operatorname{curl} - \operatorname{grad} \circ \operatorname{div}.$$

Properties

- For each triangle $\{i, j, k\}$, $\operatorname{curl} X)(i, j, k)$ measures the total flow-sum along the loop $i \to j \to k \to i$.
- If W is $\{0, 1\}$-valued edge indicator, then
  - $\operatorname{div} X(i)$ measures the inflow-outflow sum at i,
  - $\operatorname{div} \circ \operatorname{grad}$ is vertex Laplacian, $\operatorname{curl} \circ \operatorname{curl}^*$ is edge Laplacian. in all pairwise comparisons.

Theorem (Helmholtz decomposition)

Let $G = (V, E)$ be an undirected, unweighted graph and Δ_1 its Helmholtzian. The space of edge flows on G, i.e. $L^2(E)$, admits an orthogonal decomposition

$$L^2(E) = \operatorname{im}(\operatorname{grad}) \oplus \ker(\Delta_1) \oplus \operatorname{im}(\operatorname{curl}^*).$$

Furthermore, $\ker(\Delta_1) = \ker(\delta_1) \cap \ker(\delta_0^*) = \ker(\operatorname{curl}) \cap \ker(\operatorname{div})$.

Helmholtz decomposition

- **Vector calculus:** vector fields on may be resolved into irrotational (curl-free) and solenoidal (divergence-free) and harmonic component vector fields
$$\mathbf{F} = -\nabla\varphi + \nabla \times \mathbf{A} + H,$$
φ scalar potential, $\mathbf{A}$ vector potential.
- **Linear algebra:** additive orthogonal decomposition of a skew-symmetric matrix into three skew-symmetric matrices
$$X = X_1 + X_2 + X_3$$
$X_1 = se^\top - es^\top$, $X_2(i,j) + X_2(j,k) + X_2(k,i) = 0$.
- **Graph theory:** orthogonal decomposition of network flows into acyclic and cyclic components.

Hodge theory: matrix theoretic

A skew-symmetric matrix X associated with G can be decomposed uniquely

$$X = X_1 + X_2 + X_3$$

where

- X_1 satisfies
  - '*integrable*': $X_1(i,j) = s_j - s_i$ for some $s : V \to \mathbb{R}$.
- X_2 satisfies
  - '*curl free*': $X_2(i,j) + X_2(j,k) + X_2(k,i) = 0$ for all (i,j,k) 3-clique;
  - '*divergence free*': $\sum_{j:(i,j)\in E} X_2(i,j) = 0$
- $X_3 \perp X_1$ and $X_3 \perp X_2$.

Hodge theory: graph theoretic

Orthogonal decomposition of network flows on G into

$$\text{gradient flow} + \text{globally cyclic} + \text{locally cyclic}$$

where the first two components make up transitive component and

- gradient flow is integrable to give a global ranking
- example (b) is locally (triangularly) acyclic, but cyclic on large scale
- example (a) is locally (triangularly) cyclic

Rank aggregation problem revisited

Recall our formulation

$$\min_{X\in\mathcal{M}_G} \|X-\bar{Y}\|_{2,w}^2 = \min_{X\in\mathcal{M}_G}\Big[\sum\nolimits_{\{i,j\}\in E} w_{ij}(X_{ij}-\bar{Y}_{ij})^2\Big].$$

The exact case is:

Problem

Does there exist a global ranking function, $s : V \to \mathbb{R}$, such that

$$X_{ij} = s_j - s_i =: (\operatorname{grad} s)(i,j)?$$

Equivalently, does there exists a scalar field $s : V \to \mathbb{R}$ whose gradient field gives the flow X? i.e. is X **integrable**?

Answer: not always!

Multivariate calculus: there are non-integrable vector fields; cf. the film *A Beautiful Mind*:

$$A = \{F : \mathbb{R}^3 \setminus X \to \mathbb{R}^3 \mid F \text{ smooth}\}, \quad B = \{F = \nabla g\},$$
$$\dim(A/B) = ?$$

Similarly here,

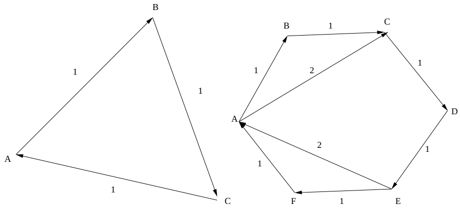

Figure: No global ranking s gives $X_{ij} = s_j - s_i$: (a) triangular cyclic, note $X_{AB} + X_{BC} + X_{CA} \neq 0$; (b) it contains a 4-node cyclic flow $A \to C \to D \to E \to A$, note on a 3-clique $\{A, B, C\}$ (also $\{A, E, F\}$), $w_{AB} + w_{BC} + w_{CA} = 0$

Boundary of a boundary is empty

Fundamental tenet of topology: (co)boundary of (co)boundary is null.

$$\text{Global} \xrightarrow{\text{grad}} \text{Pairwise} \xrightarrow{\text{curl}} \text{Triplewise}$$

and so

$$\text{Global} \xleftarrow{\text{grad}^*(=:-\text{div})} \text{Pairwise} \xleftarrow{\text{curl}^*} \text{Triplewise}.$$

We have

$$\text{curl} \circ \text{grad} = 0, \qquad \text{div} \circ \text{curl}^* = 0.$$

This implies

- global rankings are transitive/consistent,
- no need to consider rankings beyond triplewise.

Illustration

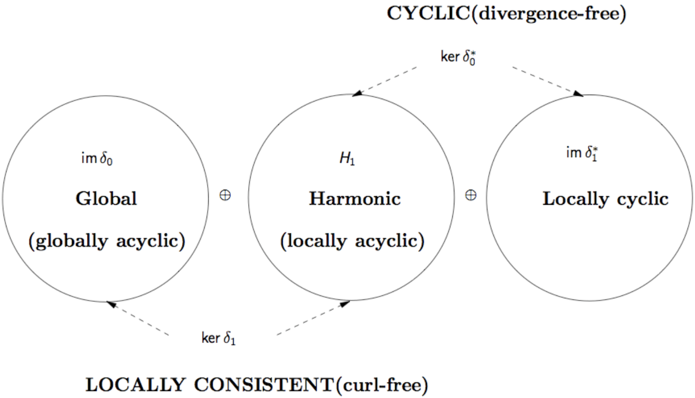

Figure: Hodge decomposition for pairwise rankings

Harmonic rankings: locally consistent but globally inconsistent

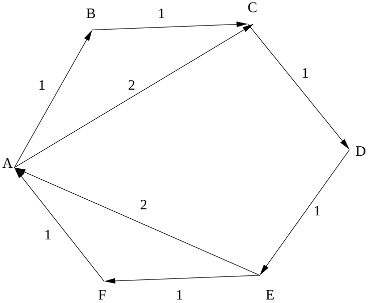

Figure: A locally consistent but globally cyclic harmonic ranking.

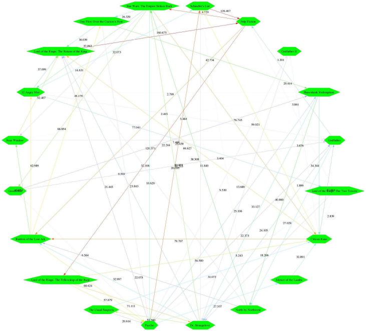

Figure: A harmonic ranking from truncated Netflix movie-movie network

Ranking interpretation of Helmholtz decomposition

- im(grad) denotes the subspace of pairwise rankings that are the gradient of score functions. Given any pairwise ranking from this subspace, we may determine a score function on the alternatives that is unique up to an additive constant.
- ker(div) $\operatorname{div} X(i) = 0$ implies alternative i is preference-neutral in all pairwise comparisons. Such pairwise rankings may be regarded as *cyclic* rankings, i.e. rankings of the form $i \succeq j \succeq k \succeq \cdots \succeq i$.
- ker(curl) denotes the pairwise rankings with zero flow-sum along any triangle in T. This corresponds to *locally consistent* (i.e. triangularly consistent) pairwise rankings.
- $\ker(\Delta_1) = \ker(\mathrm{curl}) \cap \ker(\mathrm{div})$ denotes the subspace of harmonic rankings. No inconsistencies due to small loops of length 3, i.e. $i \succeq j \succeq k \succeq i$ but has inconsistencies along larger loops of lengths > 3, i.e. $a \succeq b \succeq c \succeq \cdots \succeq z \succeq a$.
- $\mathrm{im}(\mathrm{curl}^*)$ denotes the subspace of *locally cyclic* pairwise rankings that have non-zero curls along triangles.

Erdős-Rényi random graph

Heuristical justification from Erdős-Rényi random graphs

Theorem (Kahle '07)

For an Erdős-Rényi random graph $G(n,p)$ with n vertices and edges forming independently with probability p, its clique complex χ_G will have zero 1-homology almost always, except when

$$\frac{1}{n^2} \ll p \ll \frac{1}{n}.$$

Since the full Netflix movie-movie comparison graph is almost complete (0.22% missing edges), one may expect the chance of nontrivial harmonic ranking is small.

Which pairwise ranking model might be better?

Use curl distribution:

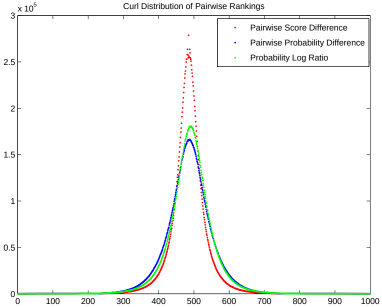

Figure: Curl distribution of three pairwise rankings, based on most popular 500 movies. The pairwise score difference in red have the thinnest tail.

Comparisons of Netflix global rankings

	Mean Score	Score Difference	Probability Difference	Logarithmic Odd Ratio
Mean Score	1.0000	0.9758	0.9731	0.9746
Score Difference		1.0000	0.9976	0.9977
Probability Difference			1.0000	0.9992
Logarithmic Odd Ratio				1.0000
Cyclic Residue	-	6.03%	7.16%	7.15%

Table: Kendall's Rank Correlation Coefficients between different global rankings for Netflix. Note that the pairwise score difference has the smallest relative residue.

Why pairwise ranking works for Netflix?

- Pairwise rankings are good approximations of **gradient flows** on movie-movie networks
- In fact, Netflix data in the large scale behaves like a 1-dimensional curve in high dimensional space
- To visualize this, we use a spectral embedding approach

Spectral embedding

Technique proposed by Goel, Diaconis, and Holmes.

- Map every movie to a point in S^5 by
$$\text{movie } m \to (\sqrt{p_1(m)}, \dots, \sqrt{p_5(m)})$$
where $p_k(m)$ is the probability that movie m is rated as star $k \in \{1, \dots, 5\}$. Obtain a movie-by-star matrix Y.
- Do SVD on Y, which is equivalent to do eigenvalue decomposition on the linear kernel
$$K(s,t) = \langle s,t \rangle^d, \qquad d = 1$$
- $K(s,t)$ is nonnegative, whence the first eigenvector captures the **centricity** (density) of data and the second captures a **tangent field** of the manifold.

SVD embedding

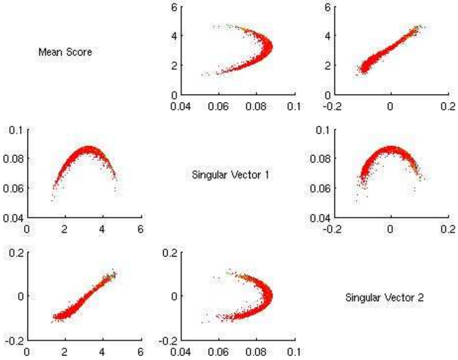

Figure: The second singular vector is monotonic to the mean score, indicating the intrinsic parameter of the horseshoe curve is driven by the mean score