An Exploration of Statistical Ranking

Wanshan Li & Yuanzhi Li & Yuze Zhou
Peking University, School of Mathematical Sciences

Contact Information:
School of Mathematical Sciences
Peking University
5 Yiheyuan Road, Beijing, China

Introduction

In this research, we went over several ranking methods based on Hodge-Rank, which targets at aggregating the pair-wise comparison data to a global ranking. First, we revise the Hodge-Rank method, which was inspired by Hodge Theory and has made great contributions in aggregate sparse and noisy comparison.

Moreover, we should be aware that the crowd is not always trustworthy, which indicates that there might be 'bad' comparisons that deviate significantly from other's decision, and thus may confuse the ranking process. Hence we adopted two methods to detect outliers, one for the comparisons and one for the annotators, both capable of dealing with incomplete and imbalanced data. These outliers should be ruled out before any ranking process are done. Finally, we did some experiments on two datasets, namely the College Ranking dataset and the Age Prediction dataset.

Hodge Rank

Mathematical Theory

Give a graph $G = (V, E)$. Denote the score of the $|V|$ vertex as s and the gradient on the edge E_{ij} as Y_{ij}. For annotators α, we have his/her estimated difference between $\hat{Y}_{ij}^{\alpha}$ and the weight ω_{ij}^{α}. Then, we can aggregate them to get gradient on edges via:

1.Weighted sum: $\hat{Y}_{ij} = \sum_\alpha \omega_{ij}^\alpha \hat{Y}_{ij}^\alpha / \sum_\alpha \omega_{ij}^\alpha \hat{Y}_{ij}^\alpha$,

2. Binary comparison: $\hat{Y}_{ij} = Pr\{\alpha|\hat{Y}_{ij}^\alpha > 0\} - Pr\{\alpha|\hat{Y}_{ij}^\alpha < 0\}$,

3. Bradley-Terry model: $\hat{Y}_{ij} = \log \frac{Pr\{\alpha|\hat{Y}_{ij}^\alpha > 0\}}{Pr\{\alpha|\hat{Y}_{ij}^\alpha < 0\}}$.

In practice, we found that Bradley-Terry model always performs better than the other two methods.

Theorem[Hodge Decomposition of Paired Ranking] Let $\hat{Y}_{ij}$ be a paired comparison flow on graph $G = (V, E)$, i.e., $\hat{Y}_{ij} = -\hat{Y}_{ji}$ for $\{i, j\} \in E$, and $\hat{Y}_{ij} = 0$ otherwise.There is a unique decomposition of $\hat{Y}$ satisfying $\hat{Y} = \hat{Y}^g + \hat{Y}^h + \hat{Y}^c$, where

$$\hat{Y}_{ij}^g = \hat{s}_i - \hat{s}_j, \text{ for some } \hat{s} \in R^V,$$
$$\hat{Y}_{ij}^h + \hat{Y}_{jk}^g + \hat{Y}_{ki}^g = 0, \text{ for each} \{i, j, k\} \in T,$$
$$\sum_{i \sim j} \omega_{ij} \hat{Y}_{ij} = 0, \text{ for each} i \in V.$$

The decomposition above is *orthogonal* under the inner product on $R^{|E|}$, $\langle u, v\rangle_\omega = \sum_{\{i,j\}\in E} \omega_{ij} u_{ij}$.

Algorithm for Hodge Rank

Algorithm 1 Procedure of Hodge Decomposition in Matlab Pseudocodes

Input A paired comparison hypergraph G provided by assessors.
Output Global score $\hat{s}$, gradient flow $\hat{Y}^g$, curl flow $\hat{Y}^c$, harmonic flow $\hat{Y}^h$
Initialization: $\hat{Y}$: numEdge-vector consisting $\hat{Y}_{ij}$ defined. W: numEdge-vector consisting ω_{ij}.

Step 1:
Compute δ_0, δ_1:// δ_0 ==gradient, δ_1=curl
$\delta_0^* = \delta_0^T * diag(W)$; //the conjugate of δ_0
$\Delta_0 = \delta_0^* \delta_0$;//Unnormalized Graph Laplacian
$div = \delta_0^* Y$;//divergence operator
$\hat{s} = lqsr(\Delta_0), div$ //global score

Step 2
Compare 1st projection on gradient flow: $\hat{Y}^g = \delta_0 * \hat{s}$;

Step 3
$\delta_1^* = \delta_1^T * diag(1./W)$;
$\Delta_1 = \delta_1 \delta_1^*$;
$curl = \delta_1 * \hat{Y}$;
$z = lqsr(\Delta_1), curl$;
Compare 3rd projection on curl flow: $\hat{Y}^c = \delta_1^* z$;

Step 4
Compare 2nd projection on harmonic flow: $\hat{Y}^h = \hat{Y} - \hat{Y}^g - \hat{Y}^c$

Robust Hodge-Rank

Outlier detection methodology

With the crowdsourcing data, there are inevitably some 'bad' comparisons, which contradict with the common-knowledge, and may lead to considerable inconsistency in aggregating global ranking.

In consideration of these potential outliers in the estimation, we adopt the Robust Hodge-Rank(RHR) method. First, we start with a linear model with respect to every single comparisons, as mentioned in the beginning,

$$Y_{ij}^\alpha = s_i - s_j + z_{ij}^\alpha,$$

where s is score like above, and z_{ij}^α is the noise term.

Since we are focusing on the outliers, here we consider a special form:

$$z_{ij}^\alpha = \gamma_{ij}^\alpha + \varepsilon_{ij}^\alpha$$

where γ_{ij} represent large magnitude deviation. In practical use, we only expect a sparse outlier component, namely $\boldsymbol{\gamma}$ should be sparse.

Hence, we adopt Huber's LASSO to gain a sparse estimation of γ:

$$\min_{\boldsymbol{s},\boldsymbol{\gamma}} \ \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{s} - \boldsymbol{\gamma}\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \quad s.t. \ \ \mathbf{1}^T \boldsymbol{s} = 1 \tag{1}$$

Note that we can separate this HLASSO (1) into two sub-problems. Denote $\boldsymbol{X}$ has singular value decomposition(SVD) $X = USV^T$, and $U = [U_1, U_2]$ where U_1 is an orthonormal basis of the column space col(X). We have:

$$\min_{\boldsymbol{\gamma}} \ \frac{1}{2}\left\|\boldsymbol{U_2^T Y} - \boldsymbol{U_2^T \gamma}\right\|_2^2 + \lambda \|\boldsymbol{\gamma}\|_1 \tag{2}$$

$$\min_{\boldsymbol{s}} \ \frac{1}{2}\left\|\boldsymbol{U_1^T X s} - \boldsymbol{U_1^T}(\boldsymbol{Y} - \hat{\boldsymbol{\gamma}})\right\|_2^2 \quad s.t. \ \ \mathbf{1}^T \boldsymbol{s} = 1 \tag{3}$$

First, we could get $\hat{\gamma}$ from (2), note that even though this estimation is somewhat biased, we can still recognize the position of those 'outliers' by (2) under mild conditions for U_2. Then, with a prior knowledge about the rough proportion of the 'outliers' (say $p\%$), we could simply choose the top $p\%$ risen in the regularization path as the 'outliers'.

After recognizing the 'outliers', we rule them out from all the comparisons to reduce bias in the estimation, compared to the combined problem (1).

Remark Note that such comparisons are based mainly on personal view, so these 'outliers' are not necessarily the untrustworthy ones.

Application on Data

We apply the hodge rank method on two data sets, the world college ranking data set and the human age prediction data set.

World College Ranking

In this dataset, we take use of the response time in outlier detection, that is to take the weights as $\sqrt{Time}$ in outlier detection (2), which reveals the haste in decision making .

Origin Rank		Deleting Outliers		Deleting Outliers and Unreliable Voters	
College	**Rank**	**College**	**Rank**	**College**	**Rank**
Princeton University, USA	1	Harvard University, USA	1	California Inst. of Tech., USA	1
Harvard University, USA	2	California Inst. of Tech., USA	2	Princeton University, USA	2
Cornell University, USA	3	Princeton University, USA	3	Yale University, USA	3
Yale University, USA	4	UC, Los Angeles, USA	4	Harvard University, USA	4
University of Cambridge, UK	5	Yale University, USA	5	University of Cambridge, UK	5
Stanford University, USA	6	Carnegie Mellon University, USA	6	Carnegie Mellon University, USA	6
UC, Los Angeles, USA	7	Cornell University, USA	7	University of Pennsylvania, USA	7
UC, Berkeley, USA	8	Stanford University, USA	8	UC, Los Angeles, USA	8
University of Oxford, UK	9	UC, San Diego, USA	9	Massachusetts Inst. of Tech., USA	9
Columbia University, USA	10	University of Cambridge, UK	10	UC, Berkeley, USA	10
Peking University, China	15	Peking University, China	23	Peking University, China	18
The University of Hong Kong	33	Tsinghua University, China	26	Tsinghua University, China	28
Hong Kong University of Science and Technology,HK	35	Hong Kong University of Science and Technology, HK	28	The University of Hong Kong	34
Tsinghua University, China	49	The University of Hong Kong,HK	29	Hong Kong University of Science and Technology, HK	37
Shanghai Jiaotong University, China	58	City University of Hong Kong, HK	67	Shanghai Jiaotong University, China	66

Table 1: Different Ranking Methods: Results

Ranking Method	Inc.Total	Inc.Har	Inc.Curl	$\|E\|$	$\|T\|$
Naive Ranking	0.880412	0.849835	0.030577	7257	28966
Robust Ranking	0.707294	0.002038	0.705256	6430	20137
FDR+Robust	0.690324	0.028571	0.661753	5286	11391

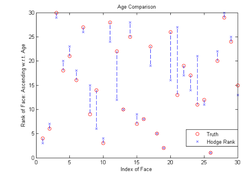
Table 2: Different Ranking Methods: Numerical features

Inconsistency

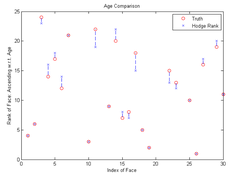
The inconsistency of the ranking results is given by$\hat{Y}^h$ and curl flow $\hat{Y}^c$. Inc.Harm($\hat{Y}$) = $\|\hat{Y}^h\|^2/\|\hat{Y}\|^2$;Inc.Curl($\hat{Y}$) = $\|\hat{Y}^h\|^2/\|\hat{Y}\|^2$. Inc.Total($\hat{Y}$) = Inc.Harm($\hat{Y}$)+Inc.Curl($\hat{Y}$)

Human Age Prediction

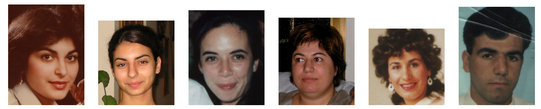
Ranking 30 faces has truth to compare. There are six evident badly-ranked people, whose photos are actually confusing. Eliminating them can give a better result as the right figure.



(a) Ranking all faces. B-T model



(b) Ranking without 6 "bad" faces



(c) 8. Age:18 (d) 9. Age:22 (e) 12. Age:32 (f) 20. Age:36 (g) 21. Age:20 (h) 24. Age:18

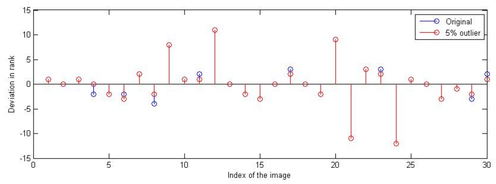


Figure 1: Deviation from the real ages

This figure shows the change in deviations after removing the outliers (taking 5%). In most cases, RHR is at least no worse than the OLS method, which verifies the validity of RHR.

Controlling FDR

In order to control FDR, the false discovery rate, it is necessary to detect the malicious annotators (such as those who deliberately choose the left side) and the strongly-biased annotators. Here, the knockoff filter based on ISS: ***Inverse Scale Space dynamics*** is chosen. The main idea is

Inverse Scale Space dynamics

First denote the model

$$Y = \delta_0 * \theta + A\gamma + \epsilon$$

where δ_0 and Y is the same as before and the (i, j) th element of A takes value 1 if the *i* the test is conducted by the *j* the annotator, otherwise, 0. The ISS algorithm gives a solution the same LASSO problem mentioned before in (1).

Denote $shrink(x) = sign(x)\ \max(|x| - 1, 0)$ below.

Algorithm 1 Procedure of ISS

Input Given parameter χ and $\triangle_t$, define k=0, $\omega^0 = 0$, $\theta^0 = (\delta_0^* \delta_0)^- \delta_0^* Y$ and $\gamma^0 = 0$. Here the $(\delta_0^* \delta_0)^-$ denotes the pseudo inverse of $\delta_0^* \delta_0$

repeat
  $\omega^{k+1} = \omega^k + A^\tau(Y - \delta_0\theta^k - A\gamma^k)\triangle_t$
  $\gamma^{k+1} = \chi * \text{shrink}(\omega^{k+1})$
  $\theta^{k+1} = \theta^k + \chi\delta_0^*(Y - \delta_0\theta^k - A\gamma^k)\triangle_t$
until $k\triangle_t > t$

Constructing knockoff features

The knockoff feature $\tilde{A}$ satisfies $\tilde{A}^\tau\tilde{A} = A^\tau A$, $A^\tau\tilde{A} = A^\tau A - diag(s)$ and $\delta_0^\tau\tilde{A} = \delta_0^\tau A$. Also s, which is a positive vector, can be constructed by semi-definite programming, which maximizes $\sum_j s_j$ and satisfies

$$0 \le s_j \le 1, diag(s) \le 2A^\tau(I - (\delta_0^\tau\delta_0)^-)A$$

Then replace A with $[A, \tilde{A}]$ and γ with $[\gamma, \tilde{\gamma}]$ in the ISS.

Generating knockoff statistics

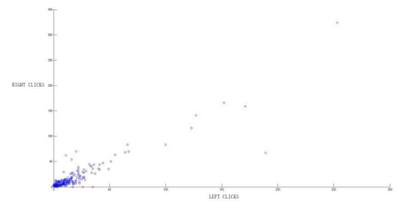
Define Z_j to be the inverse of the first entering time for A_j , i.e. $Z_j = sup(1/t : \hat{\gamma}_j \neq 0)$. Then the knockoff statistics W_j can be defined as

$$W_j = \max(Z_j, \tilde{Z}_j)sign(Z_j - \tilde{Z}_j)$$

The variables with large value of the knockoff statistics W_j are likely to be the variables to be knocked off.

Applying the knockoff method to the college ranking data

It has to be noticed that different $\tilde{A}$ may give rise to different knocked-off annotators and in order to select the knocked-off annotators more wisely, the knockoff statistics are computed for many times with different entries of $\tilde{A}$ and an averaged W_j is generated and used.



20 annotators are selected by the knockoff method. The malicious and doltish annotators are marked with red dots, but the rest are marked with blue dots. All malicious annotators who have been constantly choosing on one side are selected.

annotator	left clicks	right clicks
2684180	100	83
2687454	20	70
2687690	16	54
2687938	253	324
2691983	127	141
2707180	189	67
2931569	35	0
2931580	23	7
2934192	11	62
2959327	26	0

Table 3: Results

Although many of the annotators in the table have very strong preference on one side, some of them seem to choose on the left and right side evenly. Close analysis reveals that their choice may not make sense. For example, the annotator with id 2684180, has clicked 100 times on the left side and 83 times on the right side, which sounds very normal, but only 78 of the comparisons accords with the final ranking results, the proportion of which is comparatively very low among all annotators who have made comparisons more than 50 times.

Contribution

- Wanshan Li: Proposal of the problem, Hodge-Rank for different subsets, visualization , paper writing
- Yuanzhi Li: Proposal of the problem, Robust Hodge-Rank & outlier detection, paper writing
- Yuze Zhou: Proposal of the problem,FDR controlling , annotator selection, paper writing