Statistical Learning to Rank with Hodge Decomposition and Logistic Regression: A Case Study of World College Ranking

Dongming Huang
1000010447,Peking University

Abstract

In this paper, we will focus on the ranking methods with observation of pairwise comparison. The ranking methods include Hodge Rank, Random Walk Model, Logistic Regression and Winning Rate Ranking. We will conduct experiments based on simulation to show that Hodge Rank and Logistic Regression are desirable under the measurements of Kendall's tau Distance from the real rank and disagreement rate. Then we will apply the methods to the dataset of world college ranking.

keyword: Machine Learning, Pairwise Comparison, Hodge Theory, Logistic Regression, World College Ranking

1 Introduction

The ranking problem arises from a variety of fields such as information retrieval and recommender system[1]. The aim of a ranking model is to provide a order of some items which makes sense with the concerned dataset. This paper focuses on the pairwise comparison problem, which means that each instance (i.e. a preference of a voter) in the data is a order relation of two items, say, i and j. To represent the relation, we adopt the notation

$$i \succeq_\alpha j,$$

where the subscript α denotes the voter α.

The first challenge is to aggregate all the preferences of the voters, and the second one is to find a order r that is consistent, which means that if $i_1 \succeq_r i_2, i_2 \succeq_r i_3, \cdots, i_{m-1} \succeq_r i_m$ then we must have $i_1 \succeq_r i_m$. The solution to the first challenge depends more on the practical issue while a natural way to solve the second one is introducing a potential function $s : V \mapsto \mathbb{R}$ and defining $i \succeq_r j$ iff $s(i) > s(j)$. Here V denotes the set of items in consideration and we assume s is injection to keep things simple.

There are many methods to solve the pairwise comparison ranking problem in literature, such as Ranking SVM, RankBoost, RankNet[2][3][4]. In this paper, we study three methods including Hodge Rank[5], Random Walk Model, and Logistic Regression. In the rest of this paper, we shortly review the theory and application of Hodge Decomposition in section 2. In section 3, we give a description of the other methods to study. Some experiments based on simulation will be presented in section 4. We study the applications of these methods on the dataset of World College in section 5 and then summarize the work in section 6.

2 Hodge Rank

In the paper by Jiang-Lim-Yao-Ye[], the authors proposed the technique *HodgeRank*. The paper adopts the notation of graph G=(V,E), where V is the set of items we are interested in and edge set E is the information of comparison from the voters. If a voter α like item i more than j, then we put $Y_{ij}^\alpha = 1$; otherwise $Y_{ij}^\alpha = -1$. In the same time we let $w_{ij}^\alpha = 1$. If the voter has not give preference between items i, j, then $w_{ij}^\alpha = 0$. In short, the weight function $w : \Lambda \times V \times V \mapsto \mathbb{R}$ and for each voter $\alpha \in \Lambda$, the pairwise ranking matrix $Y^\alpha \in \mathbb{R}^{n \times n}$ is a skew-symmetric matrix, where n = #V.

Then the edge set E could be formulated as

$$E = (i, j) \in V \times V | \sum_\alpha w_{ij}^\alpha > 0.$$

If we have aggregate all the voters' preferences into a single skew-symmetric matrix $Y \in \mathbb{R}^{n \times n}$, one may hope that it could be derived from a potential function $s : V \mapsto \mathbb{R}$,

that is $Y_{ij} = s_j - s_i$ or $Y = grads$. The paper defines *combinatorial gradient operator* of a potential function with the notation *grad*. However, Y is not globally consistent generally. Let

$$\mathcal{C}_G = s : V \mapsto \mathbb{R}$$

to denote all the potential functions and let

$$\mathcal{M}_G = Im(grad)$$

to denote all the globally consistent ranking. Further more, the authors define the *combinatorial curl* operator to measure local inconsistency.

$$(curlX)_{ijk} = X_{ij} + X_{jk} + X_{ki}, \forall i, j, j, k, k, i \in E.$$

Then it is easy to see that $curl \circ grad = 0$ or $\mathcal{M}_G \subset Ker(curl)$. We denote the

Equipped with an inner product $< X, Y >_w = \sum_{ij} w_{ij} X_{ij} Y_{ij}$, the set of all the skew-symmetric pairwise functions becomes an inner product space C^1. For the set C^2 of all the skew-symmetric triplewise functions and the set C^0 of all the functions on vertices, we assign Euclid inner product. The authors prove the *Combinatorial Hodge Decomposition Theorem*, which yields the following result:

$$C^1 = Im(grad) \oplus Ker(\delta_1) \oplus Ker(curl^*),$$

where $curl^*$ is the conjugate operator of *curl* and δ_1 is the so-called *combinatorial Laplacian* (we omit the detail about it). What we are interested in is the globally consistent component that lies in $Im(grad)$.

The goal to find the $X \in \mathcal{M}_G$ best fitted to Y could be formulated as

$$\min_{X \in \mathcal{M}_G} \|X - Y\|_w^2.$$

The Hodge Decomposition Theorem tells us that the projection of Y into $Im(grad)$ is exactly the first component in the above decomposition.

To obtain the projection, we can deduce the normal equation. Suppose X is the projection of Y, then for any $Zin\mathcal{M}_G$,put $t \in \mathbb{R}$, we have

$$(\frac{d}{dt}\|X + tZ - Y\|_w^2)\Big|_{t=0} = 0,$$

which followed by

$$grad^*Y = grad^*X = grad^* \circ grads,$$

Here one can suppose $X = grads, s \in C^0$. After some arithmetic, it can be showed that

$$grad^*Y = ((W \otimes Y)^T - (W \otimes Y))\vec{1},$$

where the $\otimes$ means entrywise product,and

$$grad^* \circ grads = (\bar{D} - \bar{W})s,$$

where $\bar{W} = W + W', \bar{D} = diag(\cdots, \sum_j \bar{W}_{ij}, \cdots)$.

Given a skew-symmetric matrix Y,whose element Y_{ij} describes how good the item i is better than item j, we calculate the projection of $-Y$ into the space $\mathcal{M}_G$, denoted by $X = grads$. Then s is the potential function yield by Hodge Rank.

3 Random Walk Model and Logistic Regression

3.1 Random Walk Model

This model views the graph G as a directed graph. An edge points from j to i iff $Y_{ij} > 0$. Given a skew-symmetric matrix Y, we can remain its information by putting $\bar{Y}_{ij} = max(Y_{ij}, 0)$. Then a random walk on the graph G can be characterized by defining a stochastic matrix P as

$$P_{ij} = \frac{\bar{Y}_{ji}}{\sum_k \bar{Y}_{ki}}.$$

Here we suppose a walker begins its journey from a vertex randomly. At each step, it chooses to jump to a neighbor that is 'better' than the vertex it is at. If Y is globally consistent, then the walker must stop at the 'best' vertex since there is no 'better' neighbor. When Y is not globally consistent, it is reasonable to assume that the walk may spend more time on those vertices that are 'better' than others. So we can use the $MeanFirstAverageTime\tau$ to characterize this. Since $\tau_i = \frac{1}{\pi_i}$, we can just use the invariant distribution of the random walk instead. This idea is similar to the famous algorithm $PageRank$.

Practically, we adopt a more robust stochastic matrix $\bar{P} = \beta P + (1 - \beta)J$, where J is a stochastic matrix whose elements are constant and $\beta \in (0, 1)$ is selected to obtain robustness.

3.2 Logistic Regression

The pairwise comparison could also be formulated as a classification problem. A pair of item (i,j) can be presented in $\mathbb{R}^n$ as

$$\phi : (i, j) \mapsto (0, \cdots, \underset{i}{1}, \cdots, \underset{j}{-1}, \cdots, 0),$$

that is except the i-th element is 1 and the j-th element is -1, all element is 0. Given the k-th record of a voter's preference about (i,j), we put $y_k = 1$ if $i \succeq_\alpha j$ and $y_k = -1$ otherwise; we also put $x_k = \phi(i, j)$.

One can try to find a classification rule $f : \mathbb{R}^n \mapsto -1, 1$ which minimizes the prediction error

$$\mathbb{E}Loss(f(x_k), y_k)$$

Here the expectation is uniform distribution over the training data.

It can be observed that $\mathbb{R}^{n-1}$ is enough for the presentation of pair (i,j), because no information will lose if we remove a certain dimension. We still suppose the global rank can be derived from a potential function s. Without lose of generality, we put $s_n = 0$ and choose a particular form to model the posterior probability:

$$\mathbb{P}(y = 1|x = \phi(i, j)) = 1/(1 + exp(s_j - s_i)).$$

At the same time, the loss function is chose to be the minus logarithm of the likelihood.

Note that $< s, \phi(i, j) >= s_i - s_j$, it is easily to see that the vector s is actually the weight acquired from logistic regression model that has no constant term. With some swapping, it is possible that all the $y_k{}'s = 1$. In this case, the program is simply:

$$\max_s \prod_k \frac{exp(s_{i_k} - s_{j_k})}{1 + exp(s_{i_k} - s_{j_k})}.$$

4 Simulation Experiment

4.1 Setting

In this section, we simulate the vector s with n elements which are i.i.d. Uniform(0,1). We can directly get the true rank of the elements R_T according to s_i. We defined a parameter d so that $n^2 d$ is the number of records. Each record α tells us $i \succ_\alpha j$ if $s_i - s_j + \epsilon_\alpha > 0$. Here the ϵ_α are noise signal, with a variance of $SNR \times Var(s_i)$. Put $a_{ij}^\alpha = s_i - s_j + \epsilon_\alpha$, $a_{ji}^\alpha = -a_{ij}^\alpha$.

4.2 Aggregation of Preference

We choose the *Arithmetic mean of score differences* method to combine all the preference into a single skew-symmetric matrix Y and Z:

$$Y_{ij} = \frac{\#\{\alpha | i \succ_\alpha j\} - \#\{\alpha | j \succ_\alpha i\}}{\#\{\alpha | i \succ_\alpha j \ or \ j \succ_\alpha i\}},$$

$$Z_{ij} = \frac{\sum_\alpha a_{ij}^\alpha}{\#\{\alpha | i \succ_\alpha j \ or \ j \succ_\alpha i\}}.$$

4.3 Measurement of Ranking

Since we have the real rank, we can use the normalized Kendall's tau distance between the real rank and the result of the algorithms. The distance is defined as

$$KD(R,Q) = \frac{1}{C_n^2}\#\{(i,j) : i < j, (R_i < R_j and Q_i > Q_j) or (R_i > R_j and Q_i < Q_j)\}.$$

Besides, we can also use the Disagreement Rate to measure the results. We can calculate how many records are not consistent with the output ranking.

In the simulation part, we will utilize both methods to evaluate the algorithms. We repeat the program 10 times, then use box plots to compare.

4.4 Analysis of Random Walk Method

In section 3.1, we introduce the parameter β to get a robust invariant distribution of the stochastic matrix $\bar{P}$. Here we compare the effect of β.

We pick n=250,with d=0.1, and vary the value of SNR to see the performance of the Random Walk Algorithm by measuring the Kendall's tau distance and Disagreement Rate. The following figure shows the result, whose X-axis is the value of β.

It is obvious that in the case of noise-free, the larger β the better performance. If SNR is 0.2, then β should be smaller. In the extreme case when $SNR = 0.5$, smaller $\beta's$ yield better result. In conclusion, there is no uniformly best choice of β; it depends on the strength of noise.

In addition, we vary the number of records to validate this statement by putting $d = 0.5$ with SNR=0.5. The figure 3 has little difference from the bottomright figures above.

4.5 Comparing Different Ranking

In this section, we vary the noise level and the number of records to compare four ranking methods: Winning Rate Rank, Hodge Rank, Random Walk Model, and Logistic Regression. The simple idea of Winning Rate Rank is to rank the items according their winning percentage. Over this section, we fix the size of items n=250. For the Random Walk Model, we fixed $\beta = 0.95$.

First, when SNR is fixed to be 0.2, we pick d=0.05,0.1,0.2,0.4, and compare the performance. The result is showed in figure 4 and 5.

With regard to Kendall's tau Distance, Hodge Rank yields the best outcomes that are very close to zero and have small variances. The performance of Logistic Regression is a little better than Random Walk Model. The Winning Rate Rank, however simple, is worst among all the methods. As the parameter d increase, more records are accessible, and all the four methods yield better outcomes.

Figure 1: $\beta's$ Effect on Kendall's Distance

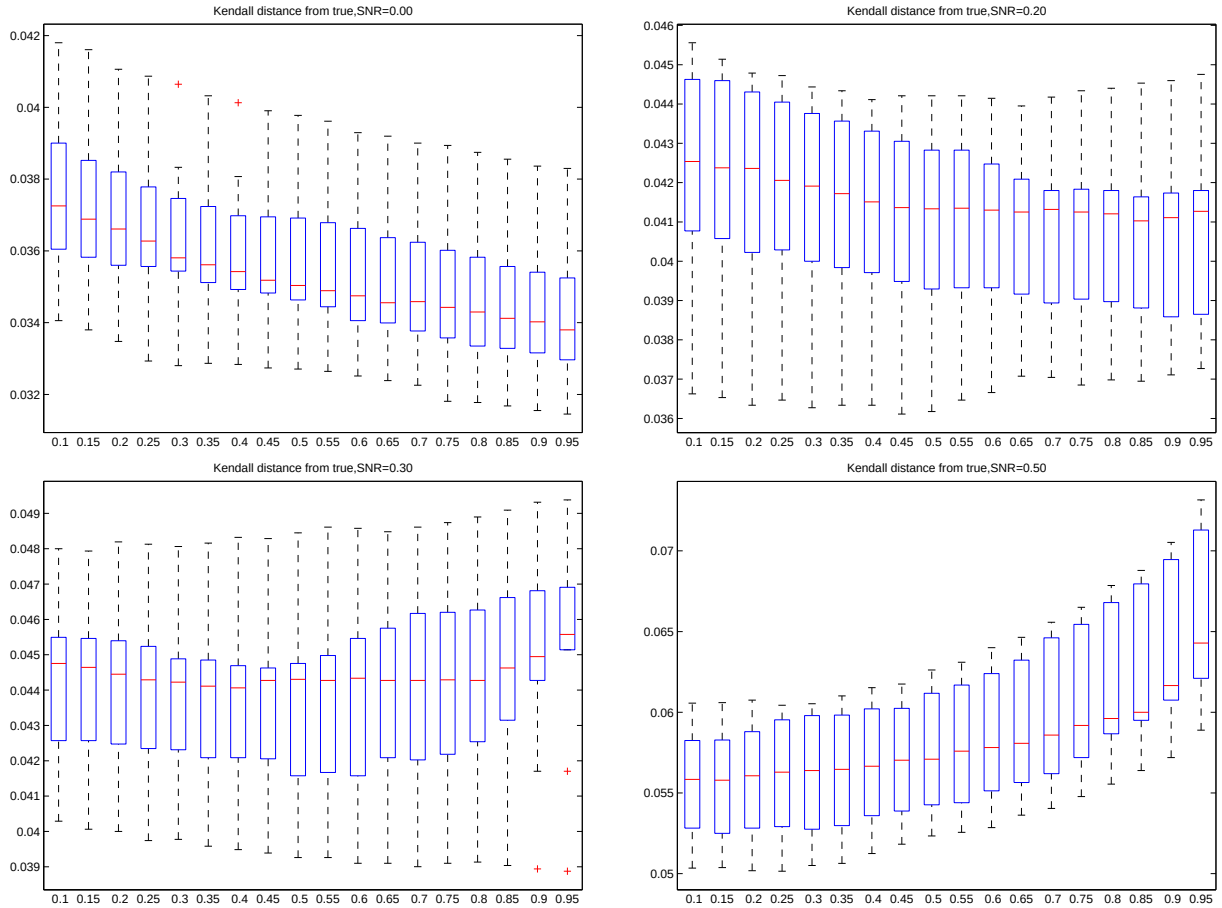


Figure 2: $\beta's$ Effect on Disagreement Rate

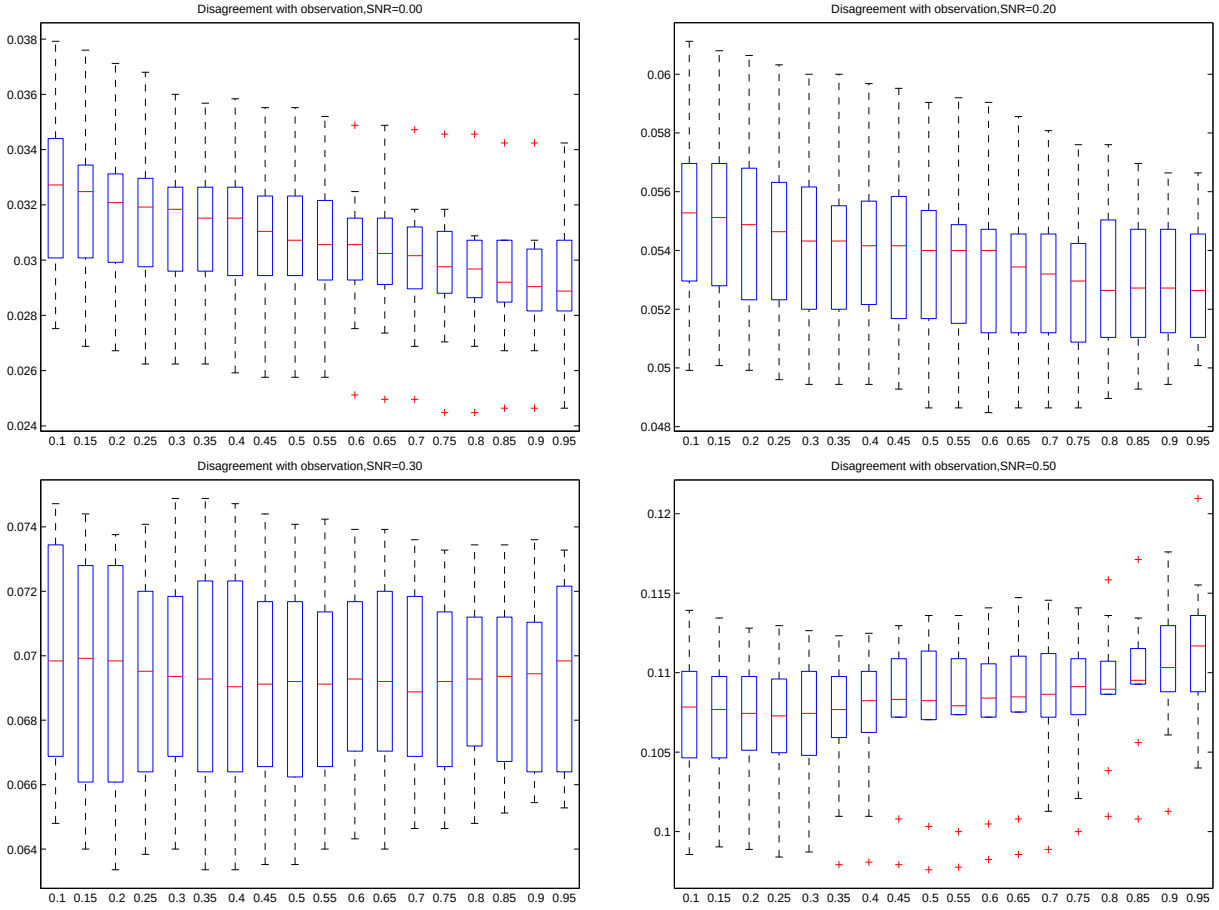


Figure 3: $\beta's$ Effect With d=0.5

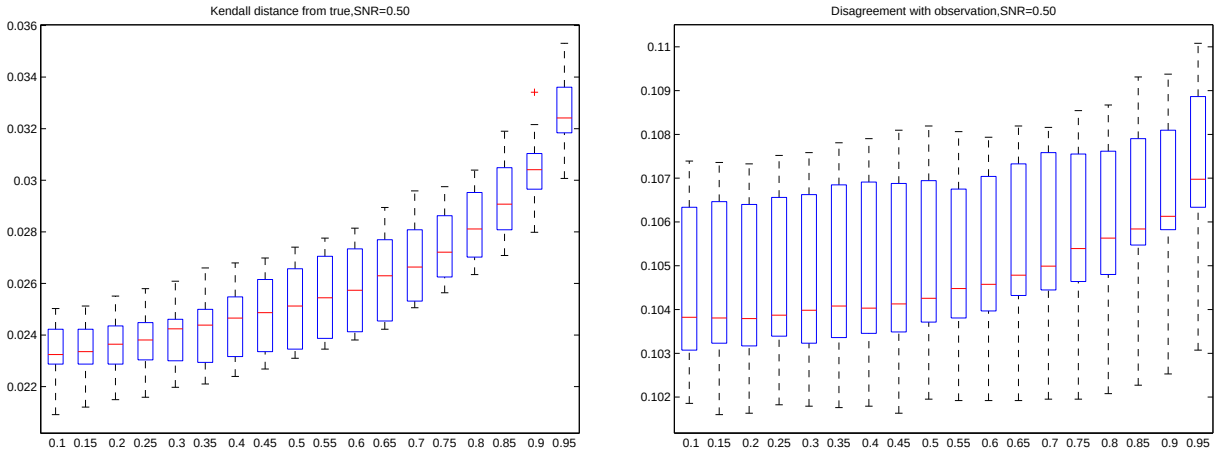


Figure 4: Kendall tau Distance From True With Different Numbers of Records

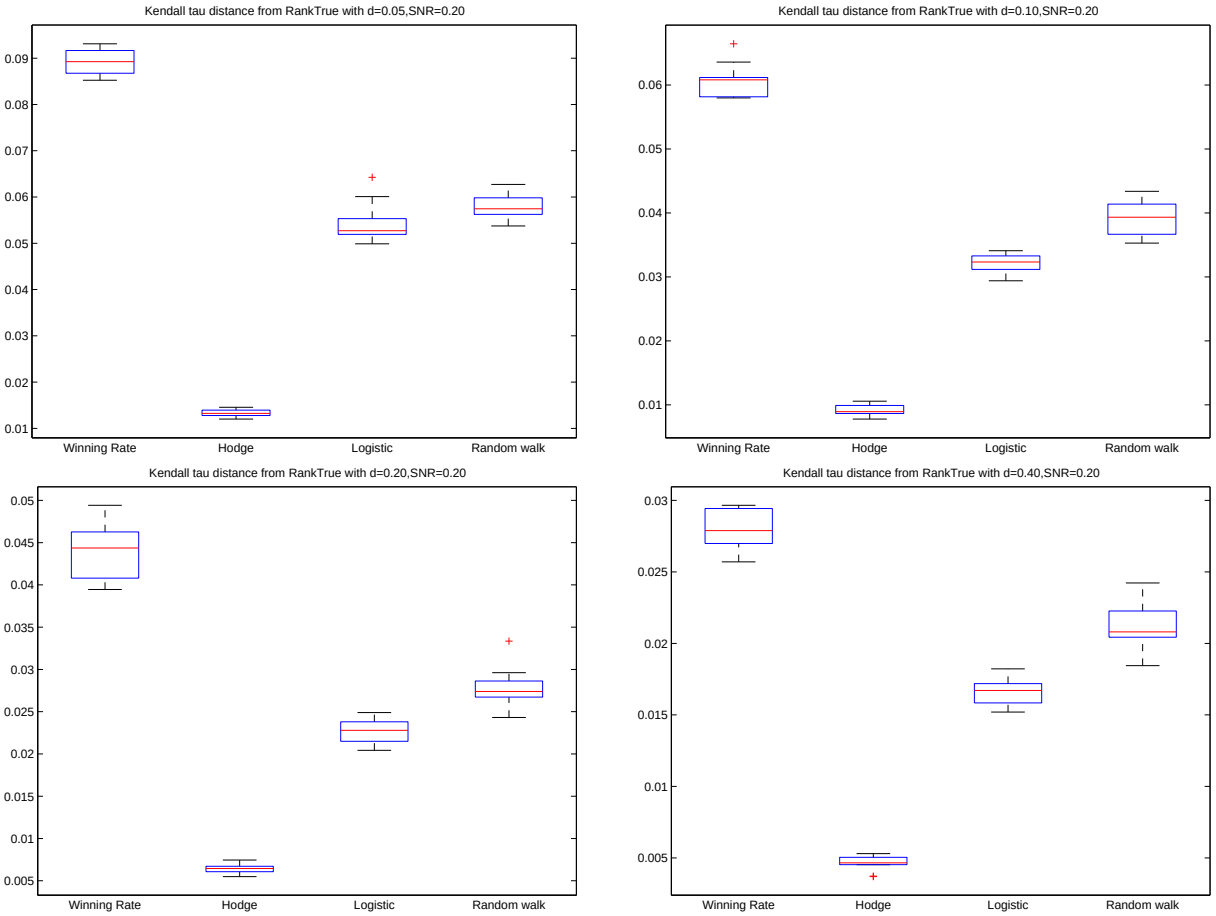
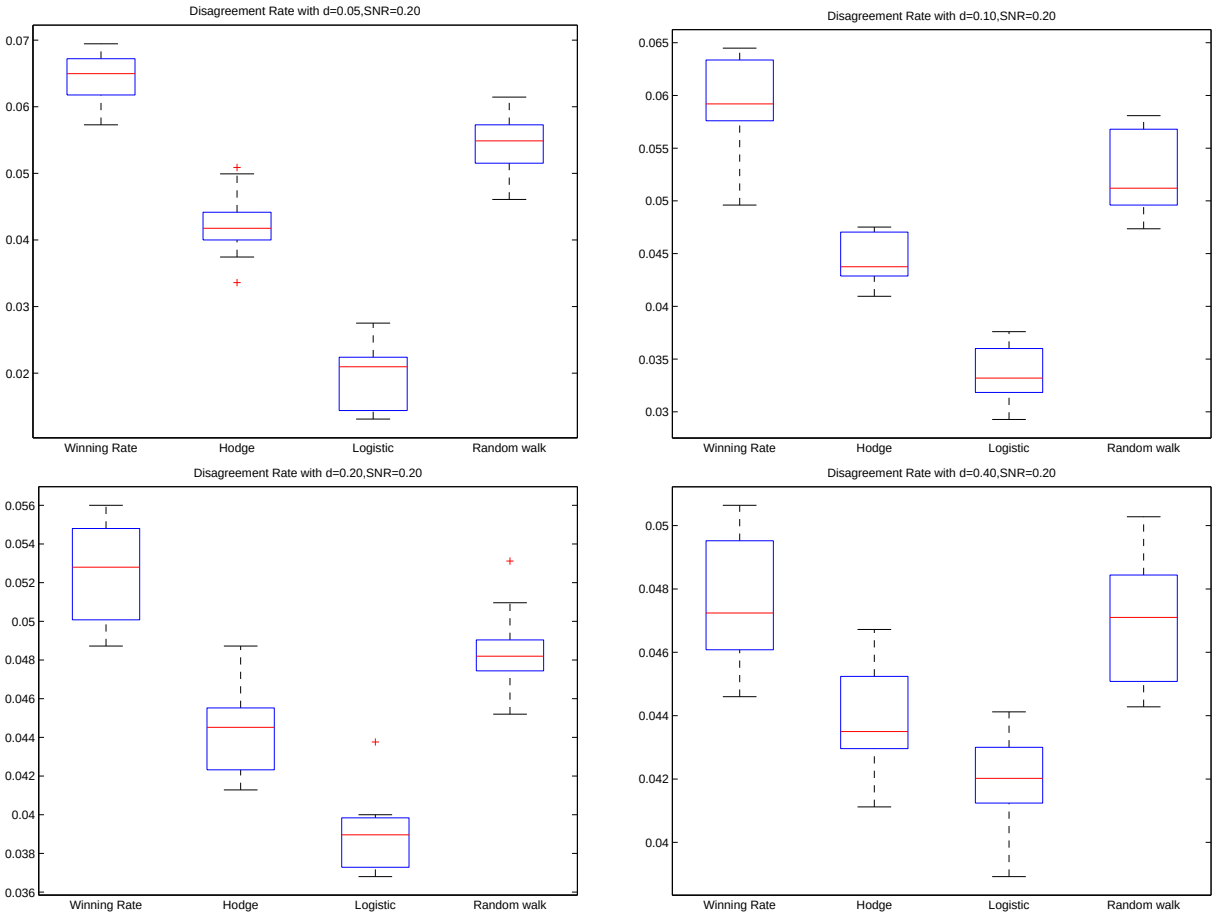


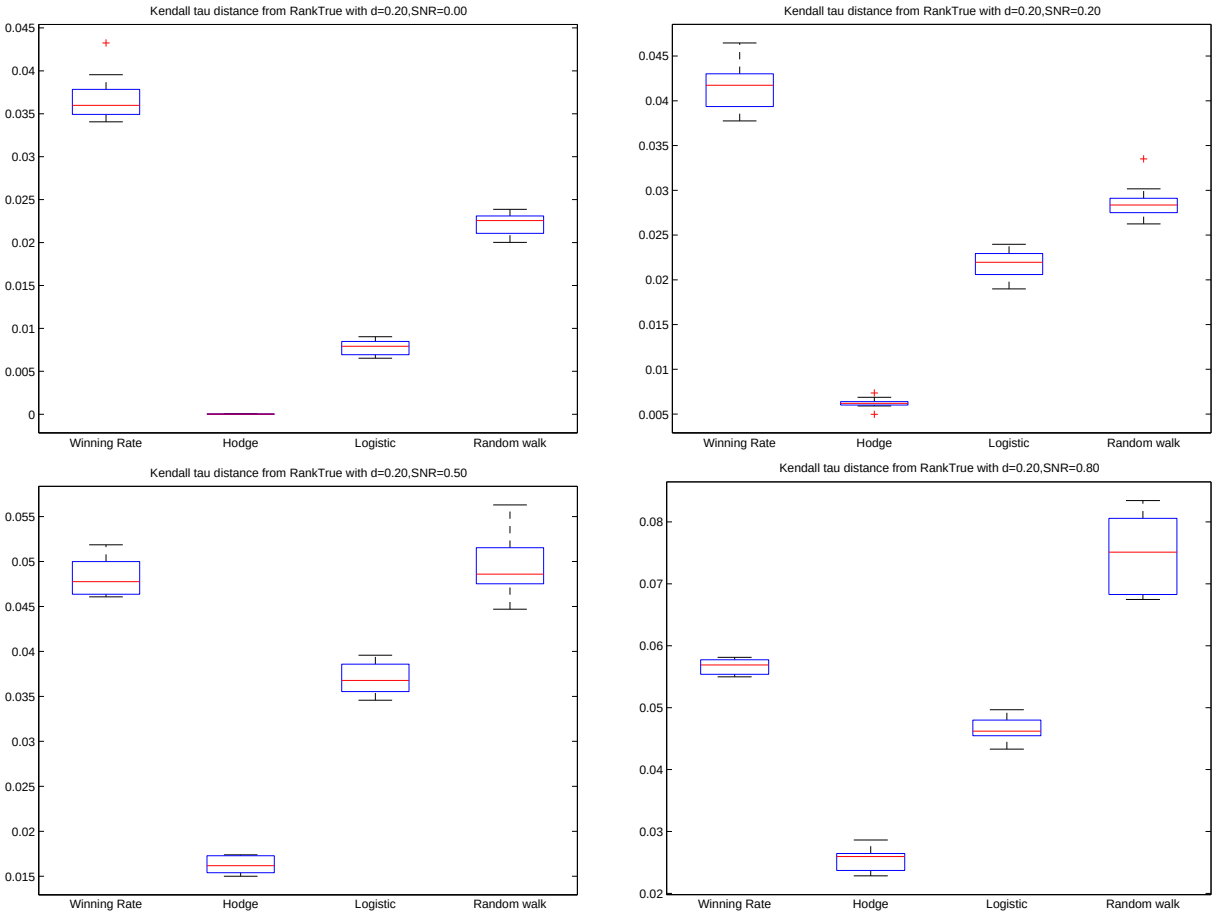
Figure 5: Disagreement Rate With Different Numbers of Records



When it comes to Disagreement Rate, Logistic Regression outperforms others. Hodge Rank becomes the second best. It is worthwhile to note that when parameter d increases, the performance of Logistic Regression is actually degenerating while those of the other three methods are improving. Among the three methods, the Winning Rate Rank improves significantly.

Then we fixed the parameter d=0.2 and pick the SNR=0,0.2,0.5,0.8. The result is showed in figure 6 and 7.

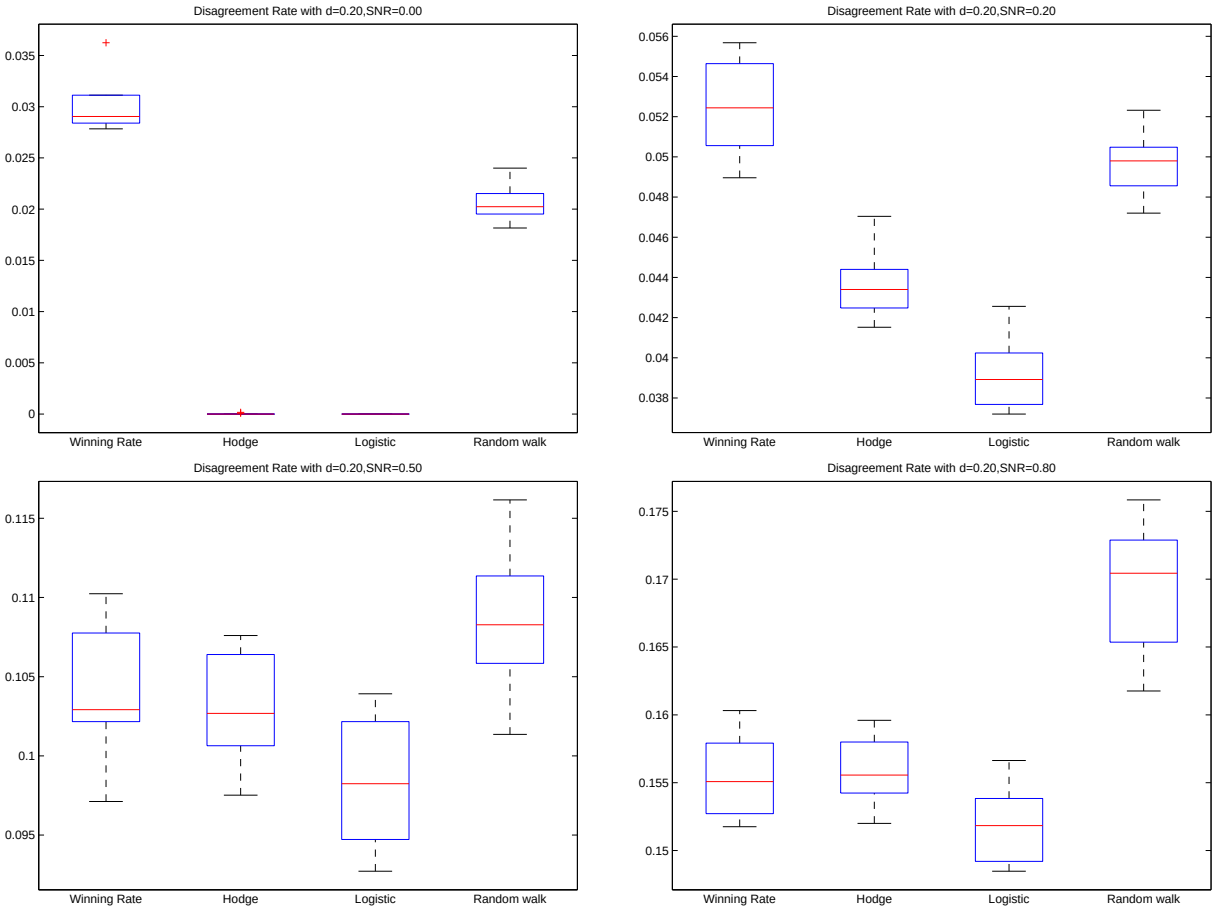
Figure 6: Kendall tau Distance From True With Different Noise Level



Again, Hodge Rank outperforms others remarkably with all noise levels. As the SNR increases, all the four methods degenerate. Random Walk Model degenerates fastest while Winning Rate Rank slowest. The sensitivity of Random Walk Model to the noise coincides with the conclusion in the last section about the effect of β. What is more, when there is a high level noise, ranking methods as simple as Winning Rate Rank make sense.

For Disagreement Rate, Hodge Rank and Logistic Regression both do extremely well in the noise-free case. In cases with noise, Logistic Regression outperforms others. As the noise level increases, the Disagreement Rate of all the methods increase. With a high level noise, Winning Rate Rank is a little better than Hodge Rank.

Figure 7: Disagreement Rate With Different Noise Level



To summarize the observation above, we can state that

1. Hodge Rank is the best algorithm among the four to recover the real order;
2. Logistic Regression is the best algorithm to minimize the Disagreement Rate because it is the goal of Logistic Regression;
3. In my experiment, Logistic Regression took far more time to yield a result;
4. Random Walk Model is so sensitive to noise that it become the last to be considered when the graph is far from consistent;
5. Simple method like Winning Rate Rank is useful in practice, especially in case with a high level of noise or a large number of voters.

5 World College Ranking

5.1 Dataset

The website http://www.allourideas.org/worldcollege ask visitors the question 'Which university would you rather attend?' when the names of two famous colleges are showed on the screen. The dataset is basically the votes by 2013-11-27. We take all the votes recorded in the file 'export_4271_votes_20131128_n5k.csv' into consideration even though some of them have a feature called 'Valid' which is 'FALSE'. We input the IDs and names of the colleges from the file 'export_4271_ideas_1384685012_f5475915b64b1e52f72423d958fce4eb133b89eb.csv', and remove the six items which have no comparison records. Then we have the number of items p=261, and number of records N=5590.

5.2 Implement and Result

We deduce the weight matrix W,voting matrix A and aggregation matrix Y from the data. There are 173 pairs of (i,j) that have been compared with different preferences.

We first apply Hodge decomposition to the matrix $-Y$, and acquired the score vector s. We can also calculate the residue $= -Y - grads$. The weighed norms of $grads$, $-Y$ and the residue are listed below.

Table 1: Norm of the Components

	Y	grad s	residue	Error Rate
$\|\bullet\|_{\mathcal{M}_G}$	102.541	55.2178	86.37	0.0236

We run the four algorithms to get four ranking results. For Random Walk Model, β is chosen to be 0.45. The Disagreement Rates are list below.

Table 2: Disagreement Rate Comparison

	Winning Rate	Hodge Rank	Logistic	Random Walk
DR	0.2823	0.2628	0.2605	0.2839

Both Hodge Rank and Logistic Regression provide us rankings that are better than the original rank on the website, at least in the sense of disagreement rate.

For particular interest, all the four methods rank Harvard University the No.1 college. Winning Rate, Hodge Rank and Logistic Regression rank Peking University the No.15 while Random Walk Model ranks PKU the No.13.

We are also interested in the similarity between each two ranks, so the Kendall's tau distances are recorded and listed below. Here we omit the normalized constant for the convenience to measure the similarity.

Table 3: Kendall's tau Distances Between Pairs of Rankings

	Winning Rate	Hodge Rank	Logistic	Random Walk
Winning Rate	0			
Hodge Rank	3887	0		
Logistic	3920	1071	0	
Random Walk	5804	4138	4071	0

From above, we see that Hodge Rank and Logistic Regression are close related to each other.

Finally, we list the Top 20 colleges according to each ranking method.

Table 4: World College Ranking

RANK	Winning Rate	Hodge Rank	Logistic Regression	Random Walk Model
1	Harvard University, USA	Harvard University, USA	Harvard University, USA	Harvard University, USA
2	California Institute of Technology, USA	Princeton University, USA	Stanford University, USA	University of California, Los Angeles, USA
3	Princeton University, USA	University of Cambridge, UK	Princeton University, USA	Carnegie Mellon University, USA
4	University of California, Berkeley, USA	University of California, Los Angeles, USA	University of Cambridge, UK	Yale University, USA
5	Massachusetts Institute of Technology, USA	Cornell University, USA	Yale University, USA	Cornell University, USA
6	Stanford University, USA	Yale University, USA	Cornell University, USA	University of California, Berkeley, USA
7	Cornell University, USA	Stanford University, USA	University of California, Los Angeles, USA	Stanford University, USA
8	University of California, Los Angeles, USA	University of Oxford, UK	University of Oxford, UK	University of Oxford, UK
9	New York University, USA	University of California, Berkeley, USA	Carnegie Mellon University, USA	Duke University, USA
10	Duke University, USA	Carnegie Mellon University, USA	University of California, Berkeley, USA	California Institute of Technology, USA
11	University of Oxford, UK	Massachusetts Institute of Technology, USA	Massachusetts Institute of Technology, USA	Columbia University, USA
12	University of Cambridge, UK	Columbia University, USA	University of California, San Diego, USA	University of Cambridge, UK
13	University of California, Santa Cruz, USA	University of California, San Diego, USA	Columbia University, USA	Peking University, China
14	University of Chicago, USA	Brown University, USA	Brown University, USA	ETH Zurich, Switzerland
15	Peking University, China	Peking University, China	Peking University, China	Boston College, USA
16	University of British Columbia, Canada	California Institute of Technology, USA	Duke University, USA	University of Michigan, USA
17	Yale University, USA	New York University, USA	Northwestern University, USA	New York University, USA
18	University of Wisconsin,Madison, USA	University of Pennsylvania, USA	University of Pennsylvania, USA	Rice University, USA
19	University of Pennsylvania, USA	Duke University, USA	University of Wisconsin,Madison, USA	University of Pennsylvania, USA
20	Brown University, USA	University of Southern California, USA	Dartmouth College, USA	University of California, San Diego, USA

6 Discussion

Based on the observation of the last two sections, we validate the statement that Hodge Rank and Logistic Regression yield more reliable rank than the Winning Rate Rank which is adopted by the website. The latter one is simple but quite useful, especially in the situation that there is a strong noise or there are many voters. The parameter β in the Random Walk Model have to be selected carefully. For the computational complexity, Logistic Regression is the slowest method.

7 Appendix

The important files in the supplement includes:

export_ideas.csv Smaller version of 'export_4271_ideas_1384685012_f5475915b64b1e52f72423d958fce4eb133b8 used to import IDs and names of the college.

export_votes.csv Smaller version of 'export_4271_votes_20131128_n5k.csv'; used to import all the pairwise comparisons of different colleges.

simulation.m Matlab code for the simulation section; used to compare four ranking methods.

simuforrandomwalk.m Matlab code for the simulation section; used to study the effect of β in the Random Walk Model.

project.m Matlab code for the case study section; used to yield four ranking of world colleges.

References

[1]Learning to rank,From Wikipedia.http://en.wikipedia.org/wiki/Learning_to_rank

[2]Herbrich R, Graepel T, Obermayer K. Large margin rank boundaries for ordinal regression[J]. Advances in neural information processing systems, 1999: 115-132.

[3]Freund Y, Iyer R, Schapire R E, et al. An efficient boosting algorithm for combining preferences[J]. The Journal of machine learning research, 2003, 4: 933-969.

[4]Burges C, Shaked T, Renshaw E, et al. Learning to rank using gradient descent[C].Proceedings of the 22nd international conference on Machine learning. ACM, 2005: 89-96.

[5]Jiang X, Lim L H, Yao Y, et al. Statistical ranking and combinatorial Hodge theory[J]. Mathematical Programming, 2011, 127(1): 203-244.