Implied Ordinal Preferences

Ian Shaw
ianedwardshaw@gmail.com

May 18, 2019

Abstract

This paper outlines a method by which to predict unexpressed preferences given only the expressed, ordinal preferences of one side of a market. With several datasets, the method is shown to be better than random. This is significant, as random assignment is the current default in the face of incomplete ordinal preferences.

[1]

[1] The code and data can be found at this github repository. https://github.com/ieshaw/Imp_Ord_Pref

Contents

1 Inspiration

This investigation was inspired by the author's previous work on developing job placement mechanisms for the Department of Defense. There was consistent difficulty in garnering complete ordinal preferences by employees for jobs.

When investigating the literature, the problem of incomplete preferences proved to arise time and time again. In 2013 Nathanson et al. review the New York Public High School choice and placements, focusing particularly on low-achieving students. [3] Their investigation notes how students' choices are "shaped and constrained by their familiarity with and proximity to specific schools" which could lead to further socioeconomic segregation. Furthermore, many assignments fall out of the top 5, requiring a second round of preference inquiry, which usually has lower participation; "The average applicant ranked between 6 and 7 schools - that is, fewer than the possible 12."

Also in 2013, Glaeser et al. review the school preferences expressed by Boston parents, noting that information is not complete within the system. [2] They mention that Boston has an open call for additional proposals to improve their assignment system.

Though there has been significant and growing work in the theoretical understanding of the phenomenon of incomplete preferences [1] [4], this investigation is intended to be a practical solution for system implementation of ordinal preference matching.

2 Formulation

2.1 Intuition

The intuition of this system is that market participants have common outlooks and drivers leading to a similarity of ranking. If we can calculate the similarity of each market participant with the others, we can use that as a coefficient to predict their ordinal preferences as a linear combination of the ordinal preferences of the others.

2.2 Mathematical Formulation

Consider a worker i and an opportunity at a firm j. Suppose there are n firms, but worker i did not express complete preferences, $n_i < n$. Suppose there are m workers in the compulsory market.

Consider the following terms

$$
\begin{aligned}
J &= \text{set of opportunities at firms in the market} \\
|J| &= n \\
P_i &\in \mathbb{Z}^{+,n\times 1} \\
P_{ij} &= \begin{cases} p & p \in \mathbb{Z}^+ > 0, \text{expressed ordinal preference of worker } i \text{ for job } j \\ 0 & \text{worker } i \text{ has not expressed an ordinal preference for job } j \end{cases} \\
P_i^C &\in \mathbb{Z}^{+,n\times 1} \\
P_{ij}^C &= \begin{cases} n-p & p \in \mathbb{Z}^+ > 0, \text{expressed ordinal preference of worker } i \text{ for job } j \\ 0 & \text{worker } i \text{ has not expressed an ordinal preference for job } j \end{cases} \\
J_i &= \{j : P_{ij} \neq 0\} \\
|J_i| &= n_i \\
J_i' &= J \setminus J_i \\
|J_i'| &= n - n_i \\
S_{i,i'} &= \text{similarity of worker } i \text{ and } i'
\end{aligned}
$$

Our proposed implied ordinal preference system P_i' is developed in the following manner. The metric r is the similarity of the worker in question with another worker multiplied by the complement ordinal ranking of that worker (so that metric increases as similar workers more prefer the positions in question), summed across all workers. The metrics are then sorted in descending order. Ties are broken randomly. Then the ordinal ranking of these metrics are considered the implied ordinal preference for the for previously unexpressed preferences.

$$
\begin{aligned}
r_{ij} &= \sum_{i'}^{m-1} S_{i,i'} P^C_{i',j} \\
R_i &= [r_{ij} : P_{ij} == 0] \\
R_i[k] &\geq R_i[k+1] \\
P'_i &\in \mathbb{Z}^{+,n\times 1} \\
P'_{i,j} &= \begin{cases} p & p \in \mathbb{Z}^+ > 0, \text{ expressed ordinal preference of worker } i \text{ for job } j \\ n_i + k & R_i[k] == r_{ij}, \text{ implied ordinal preference of worker } i \text{ for job } j \end{cases}
\end{aligned}
$$

2.3 Similarity Measures

We utilize two similarity measures in our investigation: cosine similarity and normed Euclidean distance.

Note that we use the complement of the ordinal preference, because having information up to the $n^{th} - 1$ preference is equivalent to having the n^{th} preference; we indicate having no preference information as a preference 0. Thus the most preferred choice is given a value of n.

The cosine similarity of two workers i, i' when considering their preference vectors P, is

$$S_{i,i'} = \frac{P^C_i \bullet P^C_{i'}}{||P^C_i||||P^C_{i'}||}$$

Normed Euclidean distances is calculated as

$$S_{i,i'} = 1 - \frac{||P^C_i - P^C_{i'}||}{||P^C_i||||P^C_{i'}||}$$

3 Analysis

3.1 Toy Example

Consider the incomplete preferences found in Table 1. Since most of our calculations are done with the complement of the ordinal preferences we also provide these in Table 2.

Table 1: Incomplete preferences. Blank means unexpressed.

Job Option	**Seeker 1**	**Seeker 2**	**Seeker 3**	**Seeker 4**	**Seeker 5**
Job A	1	3		1	1
Job B	2	2	1		
Job C	3	1			

Now from here we can calculate the cosine similarity of each seeker to each other, this is in Table 3.

We do a matrix multiplication of the complement of the expressed preferences with the similarity scores to get the interim preference score matrix R, found in Table 4.

Finally, we use these scores to complete the preferences in Table 5. As you can see, we have a tie to break for Seeker 3. In this case, we simply choose with random uniform probability.

3.2 Data Sources

We were able to gather ordinal preferences on job assignment with the support of the following groups:

Table 2: Complement of incomplete preferences. Unexpressed preferences are replaced with 0.

Job Option	Seeker 1	Seeker 2	Seeker 3	Seeker 4	Seeker 5
Job A	2	0	0	2	1
Job B	1	1	2	0	0
Job C	0	2	0	0	0

Table 3: Cosine Similarity between Job Seekers.

Job Seeker	Seeker 1	Seeker 2	Seeker 3	Seeker 4	Seeker 5
Seeker 1	1	$\frac{1}{5}$	$\frac{1}{\sqrt{5}}$	$\frac{2}{\sqrt{5}}$	$\frac{2}{\sqrt{5}}$
Seeker 2	$\frac{1}{5}$	1	$\frac{1}{\sqrt{5}}$	0	0
Seeker 3	$\frac{1}{\sqrt{5}}$	$\frac{1}{\sqrt{5}}$	1	0	0
Seeker 4	$\frac{2}{\sqrt{5}}$	0	0	1	1
Seeker 5	$\frac{2}{\sqrt{5}}$	0	0	1	1

Table 4: Toy Matrix R. Known preferences have blanks for scores.

Job Option	Seeker 1	Seeker 2	Seeker 3	Seeker 4	Seeker 5
Job A			$\frac{2}{\sqrt{5}}$		
Job B				$\frac{2}{\sqrt{5}}$	$\frac{2}{\sqrt{5}}$
Job C			$\frac{2}{\sqrt{5}}$	0	0

Table 5: Completed Preferences.

Job Option	Seeker 1	Seeker 2	Seeker 3	Seeker 4	Seeker 5
Job A	1	3	3	1	1
Job B	2	2	1	2	2
Job C	3	1	2	3	3

Table 6: Sources of Data Used in the Report.

Data Set	Participants	Options	% Complete
US Navy Medical Corps Doctors	59	20	34.58%
US Navy Medical Corps Hospitals	20	59	13.90%
US Navy Explosive Ordinance Displosal Officers	43	15	100%
US Navy Cryptologic Warfare Officers	28	19	100%
US Navy Cryptologic Warfare Commands	19	28	100%
US Army Officers	108	137	24.15%

3.2.1 Note on Additional Data Sources

If possible, we would be interested in extending our analysis to non-military groups; potential sources include New York and Boston School systems and the National Residency Match preference data.

3.3 Experiment Formulation

To prove the efficacy of this method, we use ordinal preference datasets that already exist, randomly dropout certain preferences, then determine how well our method predicates the artificially-unexpressed preferences. With all our datasets we run this experiment 100 times, each with 20 different percentage levels of dropout.

Consider the US Navy Cryptologic Warfare Officer dataset. There was 100% participation (complete preference coverage), with 28 participants and 19 options for a total of 532 preferences. Each experiment starts out with randomly removing 26 preferences in order to artificially drop the completeness to 95%.

To randomly dropout a preference, we randomly select a column, then remove the highest preference. If there is only one preference left in that column, we leave it and randomly select another column.

We then run our method on both similarity measures and random guessing (for baseline) to predict what the 26 dropped out preferences are. Then we take the root mean square error (RMSE) of these predicted preferences with what we know to be the true preference count.

The next dropout level is 90%, so we remove an additional 26 preferences and attempt to predict the 52 preferences that were dropped out. Then repeat the process every 5%.

This whole process is done 20 times for the 20 different dropout levels. The process then restarts, running 99 more times to accumulate 100 experiments.

3.4 Results

The plots below show the results of our experiments. As to be expected, RMSE increases with lower preference coverage because less information in the system makes preference predictions less and less reliable. Interestingly, neither similarity measure outperforms the other for all datasets. Sometimes using the cosine similarity leads to a lower RMSE than the use of Euclidean distance, other times Euclidean distance leads to better predictions. Until a compelling answer is found to as to why one similarity measure is not consistently better than the other, we recommend that any implementers of this system run this experiment on their dataset so as to decide on which similarity measure to use to complete their market's preferences.

We also include a plot of the same experiment on a dataset of random ordinal preferences. As per the intuition, since there are no underlying correlations between the columns, our method of using similarity measures performs exactly at the same rate as random guessing.

Curiously, our method is no better then random guessing, or even worse in the case of the US Army (Figure 1a) preference dataset, when preference coverage drops below 10% and especially poor when below 5%. Investigation into this phenomenon is left to future work.

4 Future Work

We hope to extend this work to non-compelling situations, where participants can express preference up to a point, after which their preference is to not participate in the market. Furthermore, a future advancement could be to normalize the preference completion to be independent of certain factors such as race or geography, which could be especially useful for the New York City Public High School match. Further research could also be supported by a better experimental error metric than RMSE. We are also curious as to why certain similarity measures preform better than others.

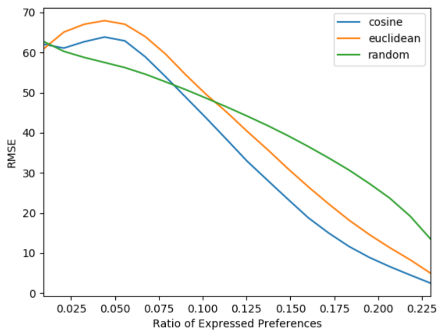
References

[1] Elena Cettolin and Arno Riedl. Revealed incomplete preferences under uncertainty: Evidence for bewley preferences. *SSRN Electronic Journal*, 01 2016.

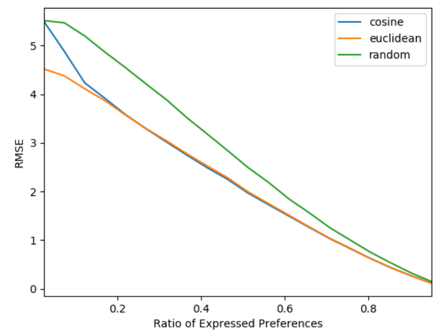
[2] S. Poftak E. Glaeser and K. Tobio. What do parents want? an exploration of school preferences expressed by boston parents. Harvard Kennedy School Faculty Research Working Paper Series, July 2013.

[3] S. Corcoran L. Nathanson and C. Baker-Smith. High school choice in new york city: A report on the school choices and placements of low-achieving students. Technical report, NYU Institute for Education and Social Policy, April 2013.

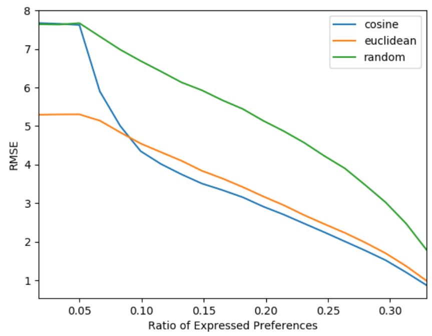
[4] Morgan McClellon. Confidence models of incomplete preferences. *Mathematical Social Sciences*, 83:30–34, September 2016.



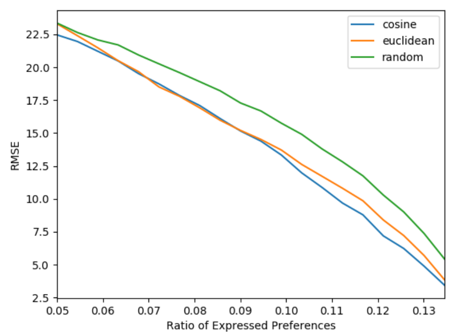
(a) Army



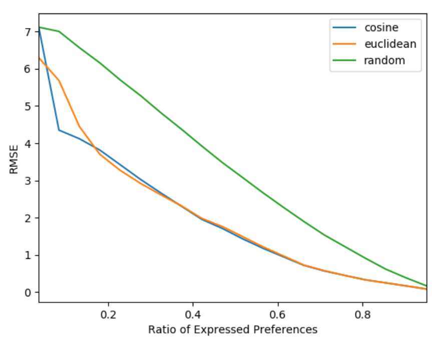
(b) Navy EOD



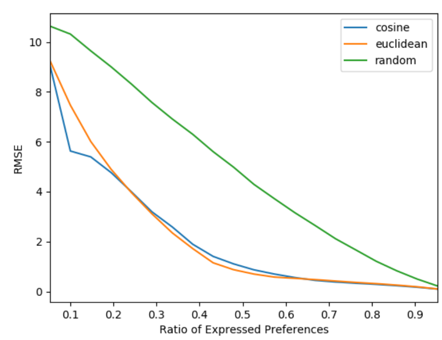
(c) Navy Medical Corps Doctors



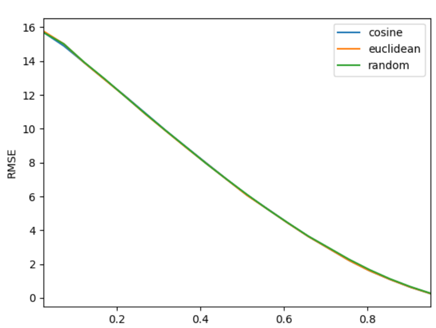
(d) Navy Hospitals



(e) Navy Cryptologic Warfare Sailors



(f) Navy Cryptologic Warfare Commands



(g) Random

Figure 1: Experiment Results