Preferences in college applications – a nonparametric Bayesian analysis of top-10 rankings

Alnur Ali
Microsoft
Redmond, WA
alnurali@microsoft.com

Thomas Brendan Murphy
School of Mathematical Sciences, UCD
Dublin 4, Ireland.
brendan.murphy@ucd.ie

Marina Meilă
Dept of Statistics, University of Washington
Seattle, WA 98195-4322
mmp@stat.washington.edu

Harr Chen
Computer Science and AI Laboratory, MIT
Cambridge, MA 02139-4307
harr@csail.mit.edu

Abstract

Applicants to degree courses in Irish colleges and universities rank up to ten degree courses from a list of over five hundred. These data provide a wealth of information concerning applicant degree choices. A Dirichlet process mixture of generalized Mallows models are used to explore data from a cohort of applicants. We find strong and diverse clusters, which in turn gains us important insights into the workings of the system. No previously tried models or analysis technique are able to model the data with comparable accuracy.

1 Introduction

Applications to degree courses[1] in Irish colleges and universities are processed by the centralized College Applications Office (CAO). When applying for college degrees, students select up to ten courses in order of preference. Courses are subsequently allocated to applicants using these preferences and the applicant's points score in an exam. Each course has a points requirement (PR) for admission; this depends heavily on the number of spaces in the course and the points scores for those who wish to do the course.

The CAO data is an example of rank data and such data arises in a number of other contexts including voting [2, 8], food marketing [12] and economic choice modeling [17]. So, the clustering of rank data is a topic of considerable interest beyond the application outlined herein, especially where the number of alternatives is large but expressed preferences are incomplete.

We study data from $N = 55737$ applications to the CAO system in the year 2000. Since every person has their own goals, utilities, and abilities, there is no complete consensus on the ranking of programs. However, we expect to find groups of applicants with similar preferences. Some clusters will be large (engineering, medicine) partly because large degree programs exist, and because of prestige and other benefits associated with certain careers. Other groups will be small because of the diversity in the population and the existence of niche careers, small programs and the geographical location of the third level institutions.

2 Model

We choose to implement a non-parametric Bayesian clustering via a Dirichlet Process Mixture (DPM) [1]. Because the data are top-t rankings, we need an appropriate statistical model for the clusters. For this we choose the Generalized Mallows (GM) model [5], that we briefly describe in this section. For a detailed presentation of the GM, the reader should consult [5, 13].

[1] Equivalently, majors, in the United States education system.

Denote by $\pi = (i_1, i_2, \ldots i_t)$ a *top-t ranking* of *length* t over a set of n items (or alternatives). In our data, $n = 533$ courses, and $t \leq 10$. Under the GM model, the probability of π is

$$GMM_{\vec{\theta},\sigma}(\pi) \;=\; e^{-\sum_{j=1}^{t} \theta_j s_j(\pi|\sigma)} / \psi(\vec{\theta}) \tag{1}$$

In the above, $\vec{\theta} = (\theta_1, \ldots \theta_t)$ are (non-negative) concentration parameters, one for each rank, while σ is the *central permutation* of the distribution, representing its mode. Note that unlike π, σ is a complete permutation of the n items. The *features* (or *codes*) $\{s_j(\pi|\sigma),\, j = 1 : t\}$ of π w.r.t σ are defined as $s_j(\pi|\sigma) = \sum_{l \succ_\pi i_j} 1_{[l \prec_\sigma i_j]}$. Thus, s_j is the number indicating one less than the rank of i_j in $\sigma \setminus \{i_{1:j-1}\}$. Finally, $\psi(\vec{\theta})$ is a tractable normalization constant that does not depend on σ [5]. The GM models are well studied and have received growing interest recently, for their interpretability and good computational properties. In particular, they have sufficient statistics [12], a conjugate prior [6], and a recently introduced algorithm for estimating DPM models via a partially collapsed Gibbs sampler [13]. We use this sampler, in conjunction with a prior which is uninformative w.r.t to the central permutation. The prior for the θ_j parameters is informative, and is described by hyper-parameters $r = [r_1\, r_2\, \ldots\, r_t]$, $r_j > 0$. We set these values to 1, which centers our θ_j priors around 0.6, a value that represents strong consensus in the clusters.

3 Experiment and statistical findings

First, we ran the DPM Gibbs sampler of [13], obtaining an ensemble of clusterings, each associated with a set of parameters for each cluster. Secondly, we use the estimated model, the original data, and additional information (about the courses, institutions, points requirements and applicant gender) to characterize the clusters and to probe their structure.

DPM clustering Because we expected many small clusters, we tuned the parameter of the DPM responsible for the prior granularity of the sampled clusterings, to a large value $\alpha = 100$. The Gibbs sampling was initialized with N singleton clusters. The number of clusters, K, after burn-in, varied between 150 and 250 clusters. Many of the clusters persisted for hundreds of Gibbs iterations; hence, for simplicity, the results presented here are obtained using the final clustering only.

Significance of ranks Each cluster c has a parameter vector $\vec{\theta}_c$ of length 10, and a central permutation σ_c of length $n = 533$. But most of these 533 ranks are noise. We designed a method to determine, for each cluster, the cutoff rank t_σ past which σ_c is noise. This cannot be determined from the candidate input lengths alone, but it can be inferred by a statistical method based on the estimated $\vec{\theta}_c$ and the cluster size, and briefly described here. We set a tail probability $\epsilon = 0.2$, and we determine τ_j, the number of items that gives $1 - \epsilon$ coverage for the exponential distribution of rank j; this is given by $\tau_j = (1/\theta_{c,j}) \ln 1/\epsilon$. For $\theta_j > 0.6$, $\tau_j = 1$, i.e. a single item will occupy rank j w.p. $1 - \epsilon$, and for $\theta_j = 0.35$, $\tau_j \approx 2$. The number $t_\sigma = \max_j j + \tau_j$ for all j that have sufficient data and $\theta_j \geq 0.1 = \theta_{min}$ is our truncation value. It represents the tail of σ_c which falls outside the $1 - \epsilon$ coverage.

We obtain a range of $t_\sigma \in [1, 12]$ for the largest 33 clusters. For these clusters, we also examined the σ_c's and $\vec{\theta}_c$ one by one, and found a remarkable agreement between our automatically determined t_σ with the courses topics and intuitive coherence of the clusters.

Results We found 33 clusters containing at least nine observations in each[2]. The cutoff point is not arbitrary: every applicant in the smallest of these clusters applied for either two or three of the home economics teaching degree courses available; this cluster would not have been found using the finite mixture models proposed in [7] or the exploratory analyses of [10, 18].

The 20 largest clusters have sizes 4500, ..., 850 and contain 92% of the data. This agrees remarkably well with the previous work of [7] which also found around 20 clusters. These clusters are strikingly *very coherent*, with all θ_j's for the first 5 ranks above 0.6 and 93% of them above 1. This again validates our model. The remaining clusters have typically even stronger consensus (which compensates for their smaller size) in spite of the parameter prior. Another phenomenon observed in these clusters is the short length of the significant part of σ_c, with t_σ values around 5. By contrast, for many large clusters, the consensus reaches all the way to θ_{10}.

[2]The remaining 131 clusters were sizes 4, 3, 3, 3, 3, 3, 2, ...

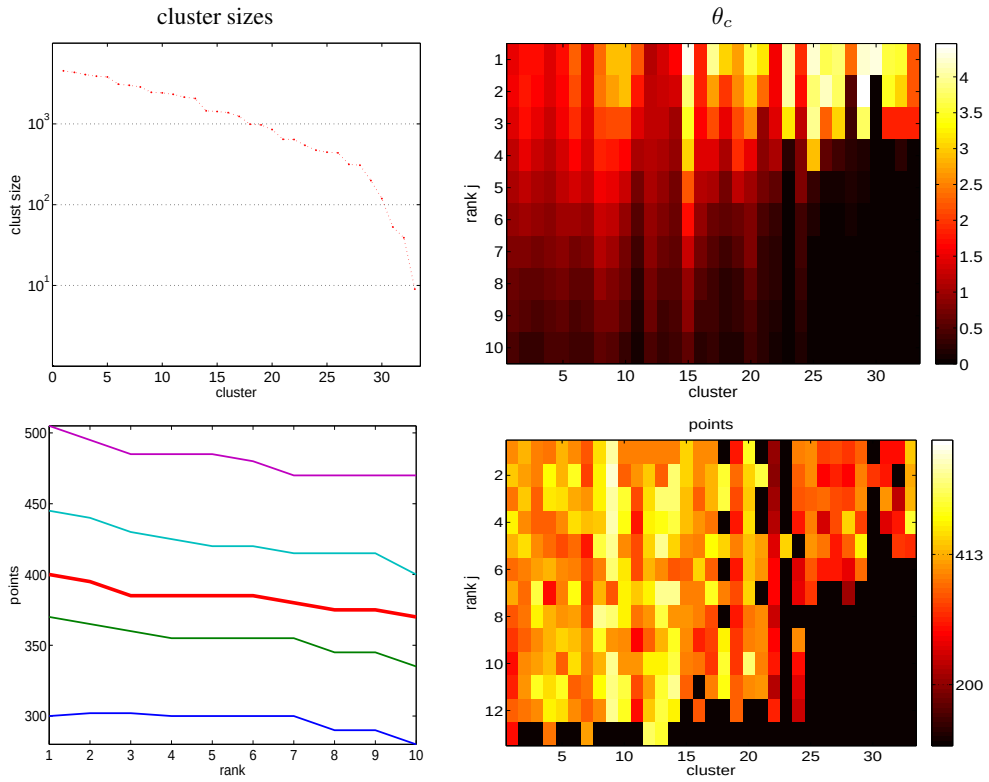

Figure 1: Right, top: heatmap for θ_j values (high values in red, low values in black) for the 33 largest clusters. Right, bottom: points requirements (PR's) for the same 33 clusters. The colorbar on the right distinguishes the median and the minimum of the PR's. For 19 courses out of 533 the PR's are missing and shown as 0 (black). Left, top: cluster size for each of the same 33 clusters. Left, bottom: quantiles (10% in blue, 30%, 50%, 70%, 90% in purple) of the PR's at each rank of the (entire) applicant data.

Table 1: Summary of the structure of the clusters found using the DPM.

Cluster	Size	Description	Male (%)	Points Average (SD)
1	4536	CS & Engineering	77.2	369 (41)
2	4340	Applied Business	48.5	366 (40)
3	4077	Arts & Social Science	13.1	384 (42)
4	3898	Engineering (Ex-Dublin)	85.2	374 (39)
5	3814	Business (Ex-Dublin)	41.8	394 (32)
6	3106	Cork Based	48.9	397 (33)
...	...	...	...	...
33	9	Teaching (Home Economics)	0.0	417 (4)

Thus from the statistical point of view, our method was successful in finding detailed structure in the data: very salient and strongly clustered groups, some large and some small. While other algorithms in the literature [7, 2] have been or could have been applied to these data, none was able to find so sharp a signal. This is a rather unexpected feature of this population. We have also fitted a GM model with 20 clusters by the EM algorithm, and an EBMS model [12]; the clusters we obtained, although somewhat meaningful, had very weak consensus, with most θ_j's around 0.07. The GM model accommodates the changing strength of preferences as the applicant completes their application, whereas a previous analysis [7] used a Plackett-Luce model [15] which does not accommodate this appropriately [4, 16].

Finally, examining the PR's along each σ_c (Figure 1, right, bottom), we note that these are *not monotonic* with the rank in any of the clusters. This is not visible in the unclustered data (Figure 1, left, bottom), where the PR very clearly decreases with j. Thus, clustering the data helps unravel a Simpson's paradox, whose significance we discuss in the next section.

4 What we learn about course applicants and the system

The CAO system is a subject of much debate in the Irish media and it makes the front page of most newspapers annually. A much touted concern is that applicants may not be selecting courses based on the subjects that they want to study, rather on the basis of PR (where higher PR courses may be more prestigious than lower PR ones) or other factors.

Since we already found that the data contain clusters, we can examine the truncated central rankings σ_c as "smoothed" expressions of each subpopulations preferences[3] (Table 1[4] summarizes some of the findings). The degree subject is the most strongly defining characteristic in the clusters. For example, science, business, arts, engineering and health sciences all characterize large clusters.

Another strong determinant of course choice is gender. The engineering clusters have a majority of male members, whereas course in the social sciences have a majority of females, and business courses tend to be gender balanced.

A further determinant of course choice is the geographical location. Three of the seven universities in Ireland are located in Dublin, one is just outside Dublin and three are distant from Dublin. The fourteen Institutes of Technology are geographically spread over the country and the smaller private colleges tend to be Dublin-based. A number of clusters are defined by subject area and location. The sixth largest cluster is characterized by courses in the two Cork-based institutions and other clusters are characterized by courses in Dublin, Galway, Waterford and Athlone. In all of these cases, there is considerable variation in the subject areas that characterize the clusters, indicating the geography is an important factor in courses selection

We do also find evidence of "prestige" as a factor, as some clusters contain a mix of high PR courses from different subject areas. Additionally, the aggregated data shows PR clearly decreasing with the rank j, as it would if the candidates were maximizing the PR. However, this hypothesis is in general completely deconstructed by the clustered data (Figure 1, right, bottom), which shows that various groups rank the courses *non-monotonically*, and strongly so, w.r.t PR. In fact, there is hardly any group where the PR is monotonic.

The $\vec{\theta}_c$ parameters for each cluster facilitate the study of the strength of preferences for applicants within each cluster and cluster coherence. In most clusters we observed that the $\vec{\theta}_c$ values are decreasing with choice levels and for the large clusters the decrease is quite slow. However, in some of the smaller clusters the $\vec{\theta}_c$ values can drop dramatically after a small number of preferences. An examination of the choices made by members of each cluster reveals a strong connection between number of choices and $\vec{\theta}_c$. The rapid decrease in $\vec{\theta}_c$ in small clusters indicates that the pool of courses of interest is less than the maximum number of preferences allowed. One cluster with 119 members is strongly characterized by students who want to study for the Evening Arts degree in UCD, which would suit people who want to study while remaining working; this is one of the only degrees that fits the needs of these applicants. This is additional evidence that factors other than simple "prestige" or PR have preponderent influence on the course choices.

5 Discussion

While we consider this analysis preliminary, the features we uncover are remarkably strong, and unlikely to be wiped by a full posterior inference (which we are currently doing, using a large subsample of our Gibbs iterations).

Thus, from the social perspective, we are satisfied to have discovered evidence that the Irish college applications system is reasonably healthy, contrary to some voiced opinions, and to have uncovered some possible factors that influence the applicants choices.

These important insights were possible using the model outlined in this paper. The flexibility of the DPM to accommodate small clusters shows that structures that are missed by other approaches can be revealed using the DPM. The clusters are easily interpretable by virtue of the fact that the

[3] We note that these preferences can be assumed to be truthful, due to the stable marriage algorithm used to make the offers.

[4] Detailed information for the 33 main clusters found by the DPM is at `http://www.stat.washington.edu/mmp/nips2010_CAO_dpmm_clusters.pdf`; information on the CAO system and PR's in 2000 is at `http://www.cao.ie`.

parameter σ_c for the GM which models each cluster is a ranking. The $\vec{\theta}_c$'s provide a tool to estimate the strength of lower ranked preferences. Finally, the careful Gibbs sampler implementation allows one to put to use this model's qualities in practice.

References

[1] Antoniak, C.E. (1974) Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Annals of Statistics*, **2** (6), 1152-1174.

[2] Busse, L.M., Orbanz, P. and Buhmann, J.M. (2007) 'Cluster analysis of heterogeneous rank data'. Proceedings of the 24th International Conference on Machine Learning. 113-120.

[3] Critchlow, D.E. (1985) 'Metric methods for analyzing partially ranked data'. Lecture Notes in Statistics, **34**, Springer.

[4] Diaconis, P. (1992) 'Foreword'. Probability Models and Statistical Analyses for Ranking Data, Fligner, M.A. and Verducci, J.S. (Eds), Lecture Notes in Statistics, **80**, Springer.

[5] Fligner, M.A. and Verducci, J.S. (1986). 'Distance based ranking models'. *Journal of the Royal Statistical Society Series B*, **48**, 359-369.

[6] Fligner, M.A. and Verducci, J.S. (1988) 'Multistage ranking models'. *Journal of the American Statistical Association*, **83** (403):892-901.

[7] Gormley, I.C. and Murphy, T.B. (2006) 'Analysis of Irish Third-Level College Applications Data'. *Journal of the Royal Statistical Society Series A*, **169** (2):361-380.

[8] Gormley, I.C. and Murphy, T.B. (2008) 'Exploring Voting Blocs within the Irish Electorate: A Mixture Modelling Approach'. *Journal of the American Statistical Association*, **103** (483):1014-1027.

[9] Huang, J. and Guestrin, C. (2010) 'Learning hierarchical riffle independent groupings from rankings.' Proceedings of the 27th International Conference on Machine Learning.

[10] McNicholas, P.D. (2006/7) 'Association rule analysis of CAO data' *Journal of the Statistical and Social Inquiry Society of Ireland*, **XXXVI**, 44-83.

[11] Marden, J.I. (1995) 'Analyzing and modeling rank data'. Chapman & Hall.

[12] Meilă, M.P. and Bao, L. (2008) 'Estimation and clustering with infinite rankings.' Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence. 393-402.

[13] Meilă, M.P. and Chen, H. (2010) 'Dirichlet Process Mixtures of Generalized Mallows Models.' Proceedings of the 26th Conference in Uncertainty in Artificial Intelligence.

[14] Neal, R.M. (2000). Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *Journal of Computational and Graphical Statistics*, **9**, 249-265.

[15] Plackett, R. (1975) The analysis of permutations. *Journal of the Royal Statistical Society Series C*, **24** (2), 193-202.

[16] Rosen, B. (1972) Asymptotic theory for successive sampling with varying probabilities without replacement, I. *Annals of Statistics*, **43** (2), 373-397.

[17] Train, K. (2003) *Discrete choice methods with simulation*. Cambridge University Press.

[18] Tuohy, D. (1998) 'Demand for third-level places. Interests, fields of study and the effect of the points for 1997' Commission on the Points System Research Paper No. 1, The Stationary Office, Dublin.