Preferences in college applications
A non-parametric Bayesian analysis of top-10 rankings

Alnur Ali\textsuperscript{1} Thomas Brendan Murphy\textsuperscript{2} Marina Meil\text{"a}\textsuperscript{3}
Harr Chen\textsuperscript{4}

\textsuperscript{1}Microsoft
\textsuperscript{2}University College Dublin
\textsuperscript{3}University of Washington
\textsuperscript{4}Massachusetts Institute of Technology
Outline

Introduction
  College Applications
  Goals
  Dataset

Model
  Data Coding
  Generalized Mallow’s models
  Dirichlet process mixture models
  Gibbs sampler

Findings
  General properties
  Overall trends

Conclusions
College Applications

- Irish college applicants apply through a central system administered by the College Applications Office (CAO).
- Applicants list up to ten degree courses in order of preference.
- Applicants are awarded points on the basis of their Leaving Certificate results; these determine course entry.
Goals

- It has been postulated that a number of factors influence course choices:
  - Institution & Location
  - Degree subject
  - Degree type (Specific vs. General)
  - Points Requirement
  - Gender

Do points requirements influence ranks?
Dataset

- We study the cohort of applicants to degree courses from the year 2000.
- The applications data has the following properties:
  - There were 55,737 applicants;
  - They selected from a list of 533 courses;
  - Applicants selected up to 10 courses.
Data Coding

• The data coding \((s_1, s_2, \ldots, s_t)\) of \(\pi | \sigma\) is defined by

\[
s_j + 1 = \text{rank of } \pi^{-1}(j) \text{ in } \sigma \text{ after removing } \pi^{-1}(1 : j - 1).
\]

Example, if \(\sigma = [a \ b \ c \ d]\) and \(\pi = [c \ a \ b \ d]\)

\[
\begin{align*}
\pi^{-1}(1) &= c & s_1 &= 2 \\
\pi^{-1}(2) &= a & s_2 &= 0 \\
\pi^{-1}(3) &= b & s_3 &= 0 \\
\pi^{-1}(4) &= d & s_4 &= 0
\end{align*}
\]

• Kendall’s distance is \(d_{Kendall}(\pi, \sigma) = \sum_{j=1}^{t-1} s_j\).
Generalized Mallow’s models

- Mallow’s model assumes that

\[
P(\pi | \sigma, \theta) = \frac{1}{\psi(\theta)} \exp \left( -\theta \sum_{j=1}^{t-1} s_j(\pi | \sigma) \right).
\]

- Can extend Mallow’s model to allow for varying precision in ranking

\[
P(\pi | \sigma, \vec{\theta}) = \frac{1}{\psi(\vec{\theta})} \exp \left( -\sum_{j=1}^{t-1} \theta_j s_j(\pi | \sigma) \right).
\]

- Location parameter \( \sigma \), scale parameters \((\theta_1, \ldots, \theta_{\text{max } t-1})\).
- \( \psi(\vec{\theta}) \) is a tractable normalization constant.
Dirichlet process mixture models

- $\tilde{p} \sim Dirichlet(\alpha/K, \ldots, \alpha/K)$
- $c_i \sim Multinomial(p_1, \ldots, p_K)$
- $\sigma_c, \theta_c \sim G_0 \propto P^0(\sigma, \theta; \nu, \tilde{r})$
- $\pi_i \sim GM(\pi_i | \sigma_c, \theta_c)$

- Prior: conjugate to $GM$, informative w.r.t. $\theta$
- DPMM benefits: no need to specify $K$ upfront, identifies both large and small clusters.
Gibbs sampler

1. Resample cluster assignments:
   1.1 Draw existing cluster w.p. \( \propto \frac{N_c-1}{N+\alpha-1} \text{GM}(\pi|\sigma_c, \theta_c) \) or Beta function approximation.
   1.2 Draw new cluster w.p. \( \propto \frac{\alpha}{N+\alpha-1} \frac{(n-t)!}{n!} \).

2. Resample cluster parameters:
   2.1 Draw \( \theta_c \) by slice sampling or a Beta distribution approx.
   2.2 Draw \( \sigma_c \) “stage-wise” or by a Beta function approx.

Beta approx. based sampler (Beta-Gibbs) faster than slice based sampler (Slice-Gibbs) (per iteration & overall time to convergence).
General properties of the clusterings

- The DPMM found 164 clusters.
- Thirty three of these clusters had nine or more members.

![Cluster Size Distribution](image)

- The clusters were characterized by a number of features.

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

- The precision parameters ($\theta_j$) were very high for top rankings.

- The $\theta_j$ values tended to decrease with $j$.

- In many cases, the $\theta_j$ values dropped suddenly after a particular point.

- The central ranking $\sigma$ for each cluster is of length 533; the $\theta_j$ values suggested a point to truncate the ranking.
Overall trends

- **Subject**
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (e.g., Science) and others opt for specific (e.g., Chemical Engineering).

- **Gender**
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.

- **Geography**
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.
Overall trends

- **Subject**
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (e.g., Science) and others opt for specific (e.g., Chemical Engineering).

- **Gender**
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.

- **Geography**
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.
Overall trends

- **Subject**
  - Subject matter is a key determinant of course choice.
  - The courses chosen are similar in subject area.
  - Some opt for general degrees (eg. Science) and others opt for specific (eg. Chemical Engineering).

- **Gender**
  - There is quite a difference in the percentage male/female applicants in some clusters.
  - Males tend to dominate CS/Engineering clusters.
  - Females tend to dominate social science/education clusters.

- **Geography**
  - There is evidence of the college location influencing choice.
  - The sixth largest cluster is dominated by courses from colleges in Cork (CIT and UCC).
  - There is evidence of a mix of subject matter and geography having a joint effect; the fourth largest cluster is dominated by engineering courses outside Dublin.
Points

- The points requirements for the courses in the truncated central rankings were not monotonically decreasing in any cluster.

- This suggests that points requirements are not important when students are ranking courses.
Conclusions & Lessons Learned

- The CAO system appears to be working more effectively than many suggest.
- The clusters revealed in this analysis tend to be cohesive in subject matter.
- The focus of possible improvements to the CAO system might be directed at how points are scored.
- The Generalized Mallows DPMM facilitated discovering small clusters that were missed in previous analyses.
- The model also allowed for the study of precision in rankings within clusters.
Questions?

Thanks!