Research Repository UCD

Provided by the author(s) and University College Dublin Library in accordance with publisher policies. Please cite the published version when available.

Title	Analysis of Irish third-level college applications data
Authors(s)	Gormley, Isobel Claire; Murphy, Thomas Brendan
Publication date	2006-03
Publication information	Journal of the Royal Statistical Society: Series A (Statistics in Society), 169 (2): 361-379
Publisher	Wiley
Item record/more information	http://hdl.handle.net/10197/7106
Publisher's statement	This is the author's version of the following article: Isobel Claire Gormley andThomas Brendan Murphy (2006) "Analysis of Irish third-level college applications data" Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(2) : 361-379 which has been published in final form at http://dx.doi.org/10.1111/j.1467-985X.2006.00412.x.
Publisher's version (DOI)	10.1111/j.1467-985X.2006.00412.x

Downloaded 2022-08-24T19:19:59Z

The UCD community has made this article openly available. Please share how this access benefits you. Your story matters! (@ucd_oa)

Analysis of Irish Third Level College Applications Data

Isobel Claire Gormley* and Thomas Brendan Murphy†

March 16, 2005

Abstract

The Irish college admissions system involves prospective students listing up to ten courses in order of preference on their application. Places in third level educational institutions are subsequently offered to the applicants on the basis of both their preferences and their final second level examination results.

The college applications system is a large area of public debate in Ireland. Detractors suggest the process creates artificial demand for 'high profile' courses, causing applicants to ignore their vocational callings. Supporters argue that the system is impartial and transparent.

The Irish college degree applications data from the year 2000 is analyzed using mixture models based on ranked data models to investigate the types of application behavior exhibited by college applicants.

The results of this analysis show that applicants form groups according to both the discipline and geographical location of their course choices. In addition, there is evidence of the suggested 'points race' for high profile courses. Finally, gender emerges as an influential factor when studying course choice behavior.

Keywords: Higher education; College course choice; Mixture models; Ranked data; EM algorithm.

1 Introduction

The Irish college applications system involves prospective college students ranking up to ten degree courses in order of preference prior to sitting their final second level examinations (Leaving Certificate). Applications are processed by the Central Applications Office (CAO) who deal with applications for all third level degree programs in Ireland.

*Department of Statistics, Trinity College Dublin, Ireland. E-mail: gormleyi@tcd.ie

†Department of Statistics, Trinity College Dublin, Ireland. E-mail: murphybt@tcd.ie

Typically, seven or eight subjects are taken for the Leaving Certificate examination. Once graded the best six examination results are used to produce a 'points' score; each grade A1, A2, B1, ..., F has an associated number of points.

Subsequent to examination grading, the CAO fixes a universal points requirement for each degree program. The points requirements are set so that applicants are offered a place in the highest preference course for which they have achieved the points requirement; in the case of applicants being tied for the last positions in a course, random allocation is used to choose which applicant is offered a place.

It is worth emphasizing that applicants do not know the required course points requirement prior to completing their application or to taking their examinations. The points requirement is influenced by the examination results of applicants who applied for the course and by the number of available positions in the course.

Some courses have minimum entry standards, for example, a sufficient standard of mathematics may be required for an engineering degree. However, the actual subjects taken at Leaving Certificate level does not have an effect on the applicants points score nor do interviews or previous examination performance; a few courses have interviews, but these are not common. The subjects Irish, English and Mathematics taken at Leaving Certificate level are entry requirements for Irish applicants for many courses but the remaining subjects are the student's choice. In addition, the Leaving Certificate can be taken several times without having any effect on an application.

International applicants are dealt with in a similar manner. For example, the UK A-Level results are converted into points - these are totalled and subsequently such applicants are allocated a course by the same method as Irish Leaving Certificate students.

Similar college applications systems are used in Australia, China and, to a degree, there are some similarities within the UK's UCAS system. Extensive details of the college applications system are available on the CAO web page (`http://www.cao.ie`).

The method of gaining entry to third level education, as managed by the CAO, is a much debated subject among the Irish media, students, parents and education circles. Many aspects of the CAO system appear annually as headlines in the Irish media – national front pages carry stories of fluctuating points requirements and volatile applicant numbers, particularly for the weeks surrounding the announcement of who is admitted to each course.

Detractors suggest that applicants are influenced by the annual media hype and rank courses according to points requirements, ensuring they study a current 'high profile' course and therefore they may ignore their vocational callings. They claim artificial demand is created for courses deemed to be of high social standing. Supporters insist it is a fair system where each applicant is dealt with in a consistent and transparent manner. The supporters claim that the so called 'points race' for courses is media generated and has no significant affect on applications.

If students are actually selecting courses according to their prestige rather than by vocational callings, then there should exist groups of applicants where the discipline of their ranked courses are quite different, but the common feature of the courses is that they have high points requirements. Therefore, if the points race drives applicants' choices, then groups of applicants ranking high points requirement courses together (such as Law, Medicine, Pharmacy, Dentistry and Actuarial Science) should be present, but where the courses are from different disciplines. On the other hand, if the system does work in its intended manner, then applicants should belong to groups where the discipline of their ranked courses is consistent.

In 1997, the Minister for Education and Science set up the "Commission on the Points System" to review the current college applications system. This lead to the publication of a report (Hyland 1999) which reviews the system and makes a series of recommendations concerning its future. A series of four research reports were also published in conjunction with the commission's final report. Of particular interest is the report of Tuohy (1998) who studies the college application data using exploratory techniques; this work is the closest to our analysis. Of some interest is the report of Lynch et al. (1999) who investigates the predictive performance of the points awarded to applicants in determining overall performance in higher education. These reports received an enormous amount of coverage in the Irish media and were discussed at length by the public. The general conclusion of the exercise was although the current system is not perfect, it works very well in practice. Clancy (1995) studies the admissions (rather than applications) data for students in Irish third level institutions, but this work is closely related to our analysis.

Our analysis focuses on analyzing the set of degree course applications made through the CAO in the year 2000; there is a separate applications system for diploma and nursing courses. We use mixture models to investigate the presence of groupings in the set of applications (Section 3.1). The resulting mixture model can be used to complete a clustering of the applicants using a model-based approach. The idea of using mixture models to cluster data has been exploited with much success by Banfield & Raftery (1993) and Fraley & Raftery (2002) amongst others. A motivation for the use of mixture models when clustering data is given by Aitkin et al. (1981) in the discussion of their paper where they say "*when clustering samples from a population, no cluster method is a priori believable without a statistical model*".

The mixture models that are employed use the Plackett-Luce model for ranked data (Section 3.2). This model describes the ranking process as a sequence of choosing the next most favored course.

The model is fitted by maximum likelihood using the EM algorithm (Dempster et al. 1977). The M-step of the EM algorithm is completed using the MM algorithm (Lange et al. 2000, Hunter 2004) as an optimization technique (Section 3.4).

In order to increase the modelling flexibility of our method, we investigate using a noise component in the mixture model to allow for applications that are very different from the majority of the remaining ones.

The resulting groupings of applicants suggested by the mixture model (Section 4) reveal that applicants generally appear to be driven by their vocational interests as discipline emerges as the defining characteristic of applicant groups. The geographical position of the institution to which an applicant applies also transpires to have a significant influence on course choice (Section 4.1). Crucially, in Section 4.2, some weight is added to the CAO system detractor's arguments. A deeper analysis of the revealed groups highlights a subtle influence of the points on the applicants choices.

A separate analysis of the male and female data suggests applicants of different gender have different course choice behaviors (Section 4.3).

2 CAO Data

The data set used in this analysis was collected in the year 2000 and it consists of the course choices of 53757 applicants to degree courses offered in Irish third level institutions. A total of 533 degree courses were selected by the applicants.

A college application made through the CAO allows an applicant to rank up to ten degree courses in order of their preference. Course places are subsequently offered using these ordered choices.

The gender of the applicants is also known and this information is only used in the analysis described in Section 4.3. There were 29338 female and 24419 male applicants in the year 2000.

A unique feature of these data is that we have a large number of applicants giving preferences for a large number of courses. In addition the applicants are restricted in the number of courses that they can rank.

3 Statistical Methodology

The data collected from the Irish college applications contains students from many different backgrounds and with many different interests. We model the course choices of these students using a mixture model, so that we can discover groups of students with different choice behavior.

The finite mixture model provides a sound model-based basis for making rigorous statements about the presence of groups and the structures of these groups. Statements can be based on sound statistical theory rather than being of a descriptive nature.

3.1 Mixture Models

We assume the course choices made by the CAO applicants form a sample from a heterogeneous population. This assumption is justified because of the differing backgrounds and interests of the college applicants. Mixture models appropriately model situations where data is collected from heterogeneous populations. Therefore, we propose using a mixture model to model the college applications data.

A finite mixture model assumes that the population consists of a finite collection of components (or groups). We assume that the (unknown) probability of belonging to component k is π_k. In addition, observations within component k have a probability density $f(\mathbf{x}_i|\underline{p}_k)$, where p_k are unknown parameters. Hence, the resulting model for a single observation is

$$f(\mathbf{x}_i) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i|\underline{p}_k).$$

Thus given our data $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M)$ and the assumption of a mixture model the likelihood is

$$L(\pi_1, \pi_2, \ldots, \pi_K; \underline{p}_1, \underline{p}_2, \ldots, \underline{p}_K|\mathbf{x}) = \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i|\underline{p}_k). \tag{1}$$

Extensive reviews of mixture modelling are given by McLachlan & Peel (2000) and Titterington et al. (1985); in addition, an excellent overview of using mixture models to produce model-based methods for clustering is given by Fraley & Raftery (2002). Previous applications of mixture models for analyzing ranked data are given in Marden (1995) and Murphy & Martin (2003) amongst others.

3.2 The Plackett-Luce Model

We need to specify an appropriate density for each component of the mixture model. Each applicant's data consists of a ranking of up to ten courses. Hence, a model that is appropriate for modelling ranked data is required. Many possible models for ranked data are described in Marden (1995), Diaconis (1988) and Critchlow (1985). Multi-stage ranking models (Marden 1995, Section 5.6) have a nice interpretation in terms of sequentially choosing items in order of preference. One parsimonious multi-stage ranking model that is easy to interpret is the Plackett-Luce model (Plackett 1975). We propose using this model for each component of the mixture model.

Plackett (1975) motivated the Plackett-Luce model in terms of modelling horse races where a vector of probabilities for each horse winning is used to construct a probability model for the finishing order. Similar characteristics can be identified between horse races and the process of ranking courses; once a course has been chosen it cannot be selected again, and following a choice being made the probability of any remaining course being selected at the next stage is altered.

In the Plackett-Luce model with parameter $\underline{p} = (p_1, p_2, \ldots, p_N)$, the probability of course c_1 being ranked in first position is p_{c_1}. The probability of course c_2 being ranked second, given that course c_1 is ranked first, is $p_{c_2}/\sum_{c\neq c_1} p_c$. That is, it is equal to the probability that c_2 is ranked first when all courses except course c_1, are available for selection. The probability of course c_3 being ranked third, given that courses c_1 and c_2 are selected first and second, is $p_{c_3}/\sum_{c \notin \{c_1,c_2\}} p_c$. The process continues to give the other placing probabilities. That is, the course choices are modelled as the product of the probabilities of each chosen course being ranked first where, at each preference level, the probabilities are appropriately normalized.

In the mixture model used in this study, we let $p_{kc(i,t)}$ denote the probability of the course chosen in tth position by applicant i being selected first, given that the applicant belongs to the kth component ($t \leq n_i$). The Plackett-Luce model then suggests the probability of applicant i's ranking conditional on coming from component k is

$$\begin{aligned} \mathbf{P}\{\mathbf{x}_i | \underline{p}_k\} &= \frac{p_{kc(i,1)}}{\sum_{s=1}^{N} p_{kc(i,s)}} \cdot \frac{p_{kc(i,2)}}{\sum_{s=2}^{N} p_{kc(i,s)}} \cdots \frac{p_{kc(i,n_i)}}{\sum_{s=n_i}^{N} p_{kc(i,s)}} \\ &= \prod_{t=1}^{n_i} \frac{p_{kc_{(i,t)}}}{\sum_{s=t}^{N} p_{kc(i,s)}} \end{aligned}$$

where, for $s \geq t$, the values of $c(i,t)$ are any arbitrary ordering of the unselected courses.

3.3 The EM Algorithm

The EM algorithm (Dempster et al. 1977) is a widely used tool for obtaining maximum likelihood estimates in missing data problems; mixture models can be formulated as having the component membership of each observation as missing data. Maximization of the likelihood function is simplified by augmenting the data to include the missing membership variables. Furthermore, the EM algorithm provides estimates not only of the model parameters but also of the unknown component memberships of the observations.

We denote the complete data by $(\mathbf{x}, \mathbf{z}) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_M, \mathbf{z}_M))$, where $\mathbf{x}_i$ is applicant i's application and

$$\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK}) \qquad \forall\ i = 1, \ldots, M$$

with

$$z_{ik} = \begin{cases} 1 & \text{if applicant } i \text{ belongs to component } k \\ 0 & \text{otherwise} \end{cases}.$$

The missing data $\mathbf{z}$ can be interpreted as an indicator of component membership. On convergence of the EM algorithm, the estimated values of z_{ik} are the conditional probabilities of applicant i belonging to component k.

Hence, the complete data log-likelihood is given as

$$\sum_{i=1}^{M} \sum_{k=1}^{K} z_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)} \right\} \tag{2}$$

The EM algorithm involves two steps, an expectation step (E-step) followed by a maximization step (M-step). In the context of finite mixture models, the expectation step estimates the unknown values of z_{ik}. The maximization step then proceeds to maximize the complete data log likelihood (2) to estimate the model parameters.

The EM algorithm is an iterative technique and continually repeats the E and M-steps until convergence to stable estimates and/or a predetermined criterion is achieved. Aitken's Acceleration Criterion (McLachlan & Peel 2000, Section 2.11) was employed in this application as a convergence criterion.

Specifically, the EM algorithm proceeds as follows:

0. **Initialize:** Choose starting values for $\underline{p}_1^{(0)}, \underline{p}_2^{(0)}, \dots, \underline{p}_K^{(0)}$ and $\pi_1^{(0)}, \pi_2^{(0)}, \dots, \pi_K^{(0)}$. Let $l = 0$.

1. **E-step:** Compute the values

$$\hat{z}_{ik} = \frac{\pi_k^{(l)} f(\mathbf{x}_i | \underline{p}_k^{(l)})}{\sum_{k'=1}^{K} \pi_{k'}^{(l)} f(\mathbf{x}_i | \underline{p}_{k'}^{(l)})},$$

   where the value $\hat{z}_{ik}$ is the estimated posterior probability of observation i belonging to group k.

2. **M-step:** Maximize the function

$$\sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)} \right\}$$

   to yield new parameter estimates $\underline{p}_1^{(l+1)}, \underline{p}_2^{(l+1)}, \dots, \underline{p}_K^{(l+1)}$ and $\pi_1^{(l+1)}, \pi_2^{(l+1)}, \dots, \pi_K^{(l+1)}$. Increment l by 1.

3. **Convergence:** Repeat the E-step and M-step until convergence. The final parameter values are the maximum likelihood estimates $\hat{\underline{p}}_1, \hat{\underline{p}}_2, \dots, \hat{\underline{p}}_K$ and $\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_K$.

The E-step is relatively straightforward when fitting a mixture of Plackett-Luce models. The optimization with respect to $\pi_1, \pi_2, \dots, \pi_K$ in the M-step is also straightforward. However, optimization with respect to $\underline{p}_1, \underline{p}_2, \dots, \underline{p}_K$ in the M-step is problematic; this optimization is discussed in Section 3.4.

3.4 MM Algorithm

The M-step of the EM algorithm aims to maximize

$$Q(\mathbf{p}) = \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \left\{ \log \pi_k + \sum_{t=1}^{n_i} \log p_{kc(i,t)} - \sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)} \right\}, \tag{3}$$

where $\mathbf{p} = (\underline{p}_1, \underline{p}_2, \dots, \underline{p}_K)$.

The $\sum_{t=1}^{n_i} \log \sum_{s=t}^{N} p_{kc(i,s)}$ term makes maximization of (3) difficult. However, Lange et al. (2000) provide a summary of a method called optimization transfer using surrogate objective functions which was later renamed as the MM algorithm. The MM algorithm is a prescription for constructing optimization algorithms more so than a directly implementable algorithm.

In order to maximize the function $Q(\mathbf{p})$, the MM algorithm forms a surrogate function that minorizes the objective function. A particular minorizing function for $Q(\mathbf{p})$ is given by $q(\mathbf{p})$, which is of the form (up to a constant)

$$Q(\mathbf{p}) \quad \geq \quad q(\mathbf{p}) \quad = \quad \sum_{k=1}^{K}\sum_{i=1}^{M} \hat{z}_{ik} \log \pi_k + \sum_{k=1}^{K}\sum_{i=1}^{M}\sum_{t=1}^{n_i} \hat{z}_{ik} \left(\log p_{kc(i,t)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} p_{kc(i,s)}^{(l)}} \right), \quad (4)$$

where the $p_{kj}^{(l)}$ values are parameter estimates.

Optimizing the surrogate function $q(\mathbf{p})$ yields new parameter values $\mathbf{p}^{(l+1)}$ which give a higher value for $Q(\mathbf{p})$; that is $Q(\mathbf{p}^{(l+1)}) \geq Q(\mathbf{p}^{(l)})$. Explicit details of the steps involved in deriving and maximizing (4) are given in Appendix A.

Note that the EM algorithm itself is an MM algorithm where the maximization of the log likelihood function (2) is transferred to a surrogate function (3); the relationship between the EM and MM algorithms is discussed in Lange et al. (2000) and Hunter & Lange (2004).

3.5 Noise Component

We investigate the inclusion of a “noise” component in the mixture with

$$(p_{k1}, p_{k2}, \ldots, p_{kN}) \quad = \quad (1/N, 1/N, \ldots, 1/N).$$

This component “soaks up” observations that have low probability of belonging to the other components. The net result is that outlying observations have less of an effect on the overall results. This component is analogous to the Poisson noise component introduced in model-based clustering (Fraley and Raftery, 2002).

When choosing the appropriate number of components in the mixture model, we also investigate if a noise component should be included in the model.

3.6 Model choice

A fundamental issue in mixture modelling is the choice of an appropriate number of components. We use the Bayesian Information Criterion (BIC) (Schwarz 1978, Kass & Raftery 1995) to choose the number of components in the mixture model.

BIC is based an approximation of the logarithm of the Bayes factor for choosing the number of components (Kass & Raftery 1995). In the case of mixture models the approximation is not valid. However, BIC has been shown to give good results (Fraley & Raftery 2002) and is theoretically supported by Leroux (1992) and Keribin (1998, 2000).

The value of the BIC is calculated as

$$\text{BIC} = 2(\text{Maximized log-likelihood}) - r \log M$$

where r is the number of parameters estimated in the model.

The log likelihood will increase with the number of fitted components, as will the number of estimated parameters. Thus the BIC strikes a compromise between model fit and model complexity by penalizing for large r.

BIC was also used to determine if a noise component (Section 3.5) should be included in the mixture model. Interestingly, we found that for the CAO data, the use of BIC suggested including a noise component in the mixture.

4 Results

Mixtures of Plackett-Luce models were fitted by maximum likelihood to the CAO data with the number of components ranging from $K = 1$ to $K = 30$. In addition, the option for allowing one of the components to be a noise component (Section 3.5) was also investigated. The mixture model with the highest BIC value was chosen and the resulting model was carefully examined.

The maximum likelihood estimates of the choice probabilities $\underline{\hat{p}}_k = (\hat{p}_{k1}, \hat{p}_{k2}, \ldots, \hat{p}_{kN})$ for each component were examined and sorted into decreasing order. From these probabilities it is possible to determine which types of courses have highest probability of being selected by applicants from the component. By examining these probabilities, the component was given a summarizing label. Clearly, we would expect that the courses with high probability values will occur with high frequency in the rankings of applicants that belong to this component.

In addition, for each component, the estimated probability of applicants coming from each component $\hat{\pi}_k$ was recorded and examined.

When the full set of all CAO applicants from is examined, the BIC values suggest that a twenty-two component mixture model should be used. The selected model had a noise component as one of the components. The mixing proportions π_k describe the percentage of the population assigned to each component. Table 1 gives the resulting twenty-two components in decreasing order of their mixing proportions.

An evaluation of Table 1 verifies the argument of supporters of the CAO system — the defining characteristic of the mixture components is the common discipline of the courses with high probabilities, as opposed to courses' common entry requirements. For example, the mixture model contains a component reflecting applicants who chose engineering courses, a component describing applicants who chose educational courses and one for applicants who chose health science courses. There is no evidence, from the examination of the probabilities, of a component representing applicants who appear to apply for high status (usually high points standard) courses. The resulting components suggest that CAO applicants do follow their vocational interests when applying to Irish institutions of third level education.

Interestingly, science based applicants are very distinctly partitioned. Applicants to biological sciences, engineering, mathematical sciences and health sciences are well segregated rather than constituting a single science component.

Table 1: The names and proportions of the twenty-two components detected when the set of all applicants were analyzed.

Component Name	**Proportion**
Business & Marketing	0.08
Hospitality Management	0.08
Arts & Humanities	0.07
Biological Sciences	0.06
Business & Commerce	0.06
Communications & Media	0.06
Construction Studies	0.06
Computer Science (Ex-Dublin)	0.05
Social Science	0.05
Munster Based Courses	0.05
Computer Science (Dublin)	0.05
Engineering	0.04
Cork Based Courses	0.04
Galway & Limerick Based Courses	0.04
Education	0.03
Health Sciences	0.03
Art & Design	0.03
Law	0.03
Mathematical & Physical Sciences	0.03
Business & Languages	0.03
Music	0.02
Noise Component	0.002

Also of note are the mixing proportion values. Ranking the components in order of mixing proportions indicates more applicants have a tendency to apply for humanities and business degrees than for more science based programs.

However, the results do require further examination and discussion; this is done in Sections 4.1–4.3.

4.1 The Geographical Effect

The components reported in Table 1 reveal important traits within the population of applicants. Most obvious are the presence of a components that highlight a geographical effect on applications.

Interestingly, five of the twenty-two components identified have a geographical basis. The Munster based courses and Cork based course components are epitomized by applicants who predominantly apply to institutions situated within the province of Munster or to institutions located in County Cork, respectively. The Galway and Limerick based component emerges from similarly motivated applicants. While possibly surprising that a geographical effect would be so well defined in such a relatively small island, readers acquainted with Irish society will be familiar with such a phenomenon. Firstly, many Irish students opt to live at home during their college studies; this differs from the situation in many other countries. Also, Irish people are very parochial and show strong affinity to their home region. People from Munster, and Cork in particular, have a very strong affinity to their region and tend to avoid travelling for their studies unless the course that they wish to study is not available in the region. Galway and Limerick are the main cities on the west coast of Ireland and a similar impetus is revealed by this component.

Also of note with regard to the geographical effect is the frequent distinction between sets of applicants who apply for degrees of similar discipline but are deemed separate based on whether or not the institutions to which they apply are located in Dublin (the capital of Ireland). Of Ireland's 3.92 million population, 1.12 million reside in County Dublin, and 2.11 million in the province of Leinster (area around Dublin). Dublin is the center of Irish governmental, financial and business dealings. Therefore some applicants are drawn to living there, while others prefer to stay away to avoid living in a large city. This goes some way in explaining why applicants view courses of a similar type as different, based on whether the location of the institution is in Dublin or not. This effect is clear on the groups of applicants applying for computer science courses, and to a lesser extent on the applicants for business, marketing and commerce degrees.

4.2 The 'Points Race'

On the surface the components determined by our model-based clustering verify the arguments of the supporters of the CAO system. Detractors insist that applicants are influenced by media hype and by the perceived social standing of some courses (revealed through their high points requirements). Examination of the reported components and their associated parameters provides deeper insight into the behavior of the CAO applicants.

We take two approaches at examining this phenomenon. We examine those courses according to the probability of the course being chosen within a component, that is using the $\mathbf{P}\{\text{Course } j|\text{Component } k\} = p_{kj}$ values (estimated by $\hat{p}_{kj}$). We also examine the posterior probability of belonging to a component given that a particular course is chosen, that is using the $\mathbf{P}\{\text{Component } k|\text{Course } j\} \propto \pi_k p_{kj}$ values (estimated by $\hat{\pi}_k \hat{p}_{kj}$).

To demonstrate that there may actually be a points race, we examine the results for the health sciences component using the two approaches described above. The results of this deeper analysis is given in Sections 4.2.1-4.2.2.

4.2.1 Examination Of Component Parameters

Table 2 shows the 30 courses with the highest probability of selection (listed in decreasing order) given that an applicant belongs to the health sciences component (see Table 1). Table 2 illustrates how components were assigned a summarizing label – from a glance it is clear that applicants belonging to this component have high probability of choosing courses leading to a degree in the health sciences sector. Many health science degree programs have high entry requirements, due to demand, a limited supply of places and the fact that these courses attract highly achieving second level students. Medicine, pharmacy, dentistry and veterinary medicine are annually reported as degree programs with higher points requirements than other courses and the resulting careers are highly esteemed within Irish society. They also are vocationally driven careers, and thus we would expect applicants to have a tendency to apply for many courses within a discipline for which they feel that they have a vocation.

Within the top 30 courses in Table 2 four have been highlighted. Arts as offered by University College Dublin, law as offered by University College Dublin and Trinity College Dublin and engineering as offered by University College Dublin. In fact applicants are almost equally likely to rank medicinal chemistry, law or therapeutic radiography given that they belong to the health sciences group. While some would, perhaps correctly, argue that a career in law is also a vocation, it could also be argued that equally so are careers such as those in the education sector. The difference between law and education degrees, in Ireland at least, is their points requirements. Law would be considered a consistently high requirement degree, whereas an education degree would have lower points requirements. There is little evidence of health science applicants choosing education programs with high probability. Therefore, some weight has been added to the assertions of CAO detractors assertion that the CAO system influences applicants to apply for courses that are prestigious (in terms of points). Another explanation is that the applicants are attracted to courses that tend to lead to high salaried professions. In any case, this implies that courses are being chosen by their status in society rather than by the discipline. How otherwise would health science applicants be

Table 2: The thirty most probable courses to be ranked on an application form, given that an applicant belongs to the health sciences component. Clearly health science degrees dominate, but the presence of high status law degrees adds some weight to the argument that applicants are influenced by the "prestige" course's points requirements.

INSTITUTION	**COURSE**	**PROBABILITY**
UCD	Medicine	0.4723
TCD	Medicine	0.2413
UCG	Medicine	0.2004
UCC	Medicine	0.1219
RCSI	Medicine	0.0610
UCD	Science	0.0351
TCD	Science	0.0297
TCD	Pharmacy	0.0280
TCD	Dentistry	0.0280
UCD	Physiotherapy	0.0260
TCD	Physiotherapy	0.0241
UCC	Dentistry	0.0233
UCD	Veterinary Medicine	0.0163
RCSI	Medicine with Leaving Certificate Scholarship	0.0153
UCG	Science	0.0140
UCC	Biological & Chemical Sciences	0.0125
TCD	Human Genetics	0.0121
UCG	Biomedical Science	0.0116
DIT	Optometry	0.0104
UCD	Radiography	0.0101
TCD	Medicinal Chemistry	0.0099
UCD	Arts	0.0092
UCD	Law	0.0091
TCD	Law	0.0085
UCD	Engineering	0.0083
TCD	Therapeutic Radiography	0.0081
TCD	Psychology	0.0074
RCSI	Physiotherapy	0.0069
RCSI	Medicine with RCSI Scholarships	0.0065
UCD	Psychology	0.0059

as likely to choose law as therapeutic radiography?

Also of note is the high probability of choosing the arts degree (in University College Dublin) and engineering (in University College Dublin). As the name partially suggests, in an arts degree students study one or two subjects from a range of arts and humanities subjects. Thus the arts degree is a very general degree that provides a broad basis from which many different career paths can emerge. In fact, it is the most frequently ranked degree program amongst all CAO applicants and has relatively achievable entry requirements. It's popularity, or perhaps it's reputation as a 'fail safe' third level choice, are possible explanations of it's high choice probability within the health science component.

The inclusion of engineering as a high probability course is less clear. The required points for engineering in University College Dublin were much lower than the health science degrees in Table 2, so the points status would not appear to be a contributing factor. It is clear health science applicants select a general science degree with high probability and perhaps are then also attracted to the general science aspects of an engineering degree. The inclusion of engineering is at least in the same scientific vein as the main thread of the health science component, whereas law degrees appear to have little in common.

4.2.2 Examination of Posterior Component Membership Probabilities

An alternative approach can be taken in the analysis of the parameter estimates by examining the posterior probability of belonging to component k given that course j was selected, that is, $\mathbf{P}\{\text{Component } k|\text{Course } j\}$. Table 3 shows the twenty-five courses whose selection gives highest posterior probability of belonging to the health sciences component.

Examination of the mixture model in this way further highlights the subtle effect the points race may have on some applicants' choices. Within the top twenty-five courses that suggest high probability of belonging to the health sciences component are Mathematics and Latin and Mathematics and Psychology, both offered by Trinity College Dublin. It appears strange to have high probability of belonging to a component dominated by health sciences courses due to the selection of either of these courses. Both are part of Trinity College's version of the general arts degree – the Two Subject Moderatorship (TSM) program. In the TSM program, students choose two modules from a range of arts and humanities subjects and study them simultaneously. However, each combination is viewed as a separate course by the CAO and due to the wide range of subjects, and therefore combinations, their choice is usually quite rare leading to sparse data. Some TSM courses lead to some strange results when analyzing the CAO data due to the rarity of some course selections within the TSM program. However the inclusion here of only two of the wide range of TSM courses suggests a contributing factor other than data sparsity. These two TSM courses both include mathematics; in that particular year points requirements for TSM courses involving mathematics were at a similar level to many of the listed health science programs in Table 3. Therefore, deeper investigation of the posterior probabilities highlights again the possibility

Table 3: Twenty-five courses whose selection on a CAO application form gives highest probability of belonging to the health sciences component.

INSTITUTION	COURSE	PROBABILITY
UCD	Medicine	0.9440
TCD	Medicine	0.9392
RCSI	Medicine	0.8886
RCSI	Medicine with Leaving Certificate Scholarship	0.8813
UCG	Medicine	0.8724
RCSI	Medicine with RCSI Scholarships	0.8010
UCC	Medicine	0.7633
TCD	Dentistry	0.5391
UCC	Dentistry	0.3674
TCD	Pharmacy	0.3110
TCD	Medicinal Chemistry	0.2725
TCD	Therapeutic Radiography	0.2662
TCD	Human Genetics	0.2579
UCD	Radiography	0.2230
TCD	Physiotherapy	0.2141
RCSI	Physiotherapy	0.2035
TCD	Mathematics/Latin	0.2013
DIT	Optometry	0.1989
UCD	Physiotherapy	0.1986
UCD	Veterinary Medicine	0.1796
TCD	Mathematics/Psychology	0.1642
UCG	Biomedical Science	0.1517
UCG	Biomedical Engineering	0.1093
TCD	Science	0.1053
TCD	Occupational therapy	0.0992

of a subtle effect that a course's points requirements may have an affect on CAO applicants.

Why the focus on law programs and mathematics programs as examples of the points race? Other high points courses such as Actuarial and Financial Studies (in UCD) also appear within the top 50 programs in both views of the model; again, this course appears to be a strange course to appear amongst a component dominated by the "vocational" health science sector.

Why the focus on the health sciences component only? It seems natural to also consider the law component that is also deemed as high points and high status. The points race effect is also apparent here – psychology in both UCD and TCD, which had high entry requirements that year, have high probability of being selected given that an applicant belongs to the law component. While law and psychology have some similarities, they would not be deemed as members of the same discipline suggesting some element of the points race is present. However, examination of the posterior probabilities for the law component gives less of an indication of the presence of a points race. It seems the points status of courses has more of an effect in the health science component than in the other components in the mixture model.

4.3 The Gender Effect

The gender of each CAO applicant in 2000 was available in addition to their course choices; of the 53757 applicants, 24419 were male. The data was partitioned according to applicant gender and mixtures of Plackett-Luce models were fitted to the two resulting data sets. Examination of the resulting parameter estimates, $\hat{p}_{kj}$, led to the summarizing component labels as outlined in Table 4.

The resulting mixtures fitted to the partitioned data provide good insight into the different choice behavior of the male and female applicants. The predominant aspect of the component labels is subject discipline, thus enhancing the supporting view of the CAO that applicants are inclined to follow their vocational interests. The geographical effect discussed in Section 4.1 is again apparent, but it is more apparent in the male results. In particular, some male components reveal a common discipline but at different geographical locations; this occurs more so than in the female components. For example, the male engineering applicants are partitioned by the location of the institution in Dublin, as are the computer science applicants.

Stereotypical differences between the two genders are very apparent in the resulting components – there appears to be distinct components for females in social science, art and design, music and education whereas female applicants with an interest in engineering and computer science are grouped together. Not only are male the engineering and computer science components separate, they are further divided within these disciplines by geography. Further, the largest component (with probability 0.09) in the male results involves construction studies courses whereas this does not appear as a distinct component in the female

Table 4: The resulting 16 components from analysis for the female applicants, and the resulting 17 from the male applicants.

FEMALE RESULTS		MALE RESULTS	
Component Label	**Proportion**	**Component Label**	**Proportion**
Hospitality Management	0.11	Construction Studies	0.09
Social Science	0.11	Communications & Journalism	0.09
Business & Marketing (Dublin)	0.09	Business & Marketing (Dublin)	0.09
Biological Sciences	0.08	Computer Science (Ex-Dublin)	0.08
Cork Based Courses	0.08	Hospitality Management	0.07
Applied Computing (Ex-Dublin)	0.07	Computer Science (Dublin)	0.07
Communications & Journalism	0.07	Arts/Humanities	0.06
Business & Commerce (Ex-Dublin)	0.07	Engineering (Ex-Dublin)	0.06
Law & Psychology	0.06	Business & Commerce	0.06
Galway & Limerick Based Courses	0.06	Cork Based Courses	0.06
Education	0.05	Law & Business	0.06
Engineering & Computer Science	0.04	Engineering (Dublin)	0.05
Art/Design & Music	0.04	Sports Science & Education	0.05
Business & Languages	0.04	Science	0.04
Health Sciences	0.04	Limerick Based Courses	0.04
Noise Component	0.003	Health Sciences	0.03
		Noise Component	0.004

results.

Other results of interest are the popularity of biological sciences amongst females whereas males have a general science component in their results, both genders have education components but the male education component also has a sports aspect.

In addition, in close similarity to the results for all applicants (see Section 4.2.1), the male health sciences component contains three law degrees in the top thirty most probable courses. Similar results are revealed for the females, but the probability of selection of the law courses is lower within the health sciences component.

4.4 Clustering of Applicants

A major advantage of fitting mixture models via the EM algorithm, as detailed by Fraley & Raftery (1998), is that the value $\hat{z}_{ik}$ at convergence is an estimate of the conditional probability that observation i belongs to component k; these values can be used to cluster observations into groups. A clustering of the set of applicants is simply achieved by examining $\max_k \mathbf{P}\{\text{Component } k|\text{Application } i\}\ \forall\, i$ and assigning applicants to the group for which the maximum is achieved.

The clustering of applicants can be scrutinized in different ways. As suggested by Bensmail et al. (1997), the uncertainty associated with an applicant's component membership can be measured by $U_i = \min_{k=1,\ldots,K}(1 - \mathbf{P}\{\text{Component } k|\text{Application } i\})$. When i is very strongly associated with group k then $\mathbf{P}\{\text{Component } k|\text{Application } i\}$ will be large and so U_i will be small. Figure 1 illustrates the uncertainty associated with the clustering of the male and female applicants.

Figure 1: Uncertainty in the clustering of female and the male applicants.

Clearly, the clustering uncertainty values tend to be very small, with 61% of females and 59% of males classified with an uncertainty of less than 0.05. Summary statistics for the uncertainty values further demonstrate how well the model allocates applicants to components; these are given in Table 5.

Table 5: Summary statistics associated with the clustering uncertainty of male and female applicants.

	1st Quartile	**Mean**	**3rd Quartile**
FEMALE	0.0002	0.1228	0.1866
MALE	0.0002	0.1301	0.2043

5 Discussion

This paper presents a model-based statistical analysis of degree level applicants to Irish institutions of third level education. The methods seek to find groups of similar applicants, and to draw conclusions about the merits and failures of the centralized applications system from the defining characteristics of these groups.

A top level view of the groups of applicants suggested by the analysis verifies a supporting view of the CAO system — applicants appear to follow their vocational interests and rank their third level course choices in a manner which reflects this. The analysis suggests that the majority of CAO applicants use the system as it is intended and rank courses in view of their genuine preferences and/or career choice. However, it is apparent that more subtle influences also contribute to course choice and a detailed examination of the mixture components indicates the presence of the reported 'points race'. It appears there are those who choose courses on the points levels of previous years and therefore on the prestige attached to some of these courses.

While most discussions of the CAO system in Irish education circles focus on the influence of the 'points race' this work highlights other factors which have an influence on an applicant's course choice. The geographical location of the institution to which an applicant applies has a clear affect on the choice process. Whether this is due to a vocational desire to study a particular course in a specific institution, the desire to live in a certain area or because of financial viability, it is a striking feature of the groups of applicants. A course's geographical location appears to be almost as important as vocational interest in an applicant's choice process. Whether this feature is a benefit of the CAO system or not remains to be researched.

Further to the effects of vocation, geography and the points race, the gender of the applicant also affects course choice. Geography and the points race may have a larger effect on male applicants than on females. Stereotypical gender differences are also apparent — only 4% of female applicants are 'classified' as engineering and computer science students compared to 26% of the population of male applicants. Further differences (see Section 4.3) indicate that males and females need to be targeted in difference ways, with regard to third level education, and this should be of interest to third level institutions and to governmental education departments.

In terms of the model employed within components, the Plackett-Luce model performs well when modelling the rankings of the preferred third level choices of CAO applicants. The model does suffer from the independence from irrelevant alternatives (IIA) within a component (see Train 2003). While it can be argued that such models are unrealistic in some situations, in this application the model appears to provide a realistic representation of the course choice process. It is worth noting that the Plackett-Luce model leads to choice probabilities that satisfy Luce's Choice Axiom (see Marden 1995, Section 5.13.1).

The only covariate available for this analysis is the gender of the applicant – relationships between course choice and other covariates are very likely to be present. Expanding the analysis to include other covariates would also be desirable, but further covariates were not available for this study.

6 Acknowledgements

Both authors would like to thank Adrian Raftery and other members of the Working Group on Model-Based Clustering at the University of Washington, Seattle for important inputs into this research.

Isobel Claire Gormley was supported by a Government of Ireland Research Scholarship in Science, Engineering and Technology provided by the Irish Research Council for Science, Engineering and Technology, funded by the National Development Plan.

A MM Algorithm Calculations

We detail the steps involved in the formation of the required surrogate function to which maximization is transferred in the M-step of the EM algorithm and from which the maximum likelihood estimates $\hat{p}_{kj}$ are derived. This derivation is closely related to calculations given in Hunter (2004). The general reviews of the MM algorithm given by Lange et al. (2000) and Hunter & Lange (2004) are also of interest.

The construction of a surrogate function relies on the exploitation of properties of convex functions. For convex function $f(x)$ with differential $df(u)$, the *supporting hyperplane property of a convex function* f,

$$f(x) \geq f(y) + df(y)(x - y) \qquad x, y \geq 0 \tag{5}$$

provides a linear minorizing function that can be utilized as a surrogate function in an optimization transfer algorithm. Sometimes it is preferable to form a quadratic or higher order surrogate function. Expanding (5) using higher order expansions can yield such higher order functions.

In this application, we wish to construct a minimizing surrogate function to be iteratively maximized. This iterative maximization gives a sequence of parameter estimates with increasing values for the expected complete data log-likelihood function (3). The strict convexity of the $-\log(x)$ function implies that

$$-\log(x) \quad \geq \quad -\log(y) + 1 - \frac{x}{y}.$$

We let $f(x) = -\log \sum_{s=t}^{N} p_{kc(i,s)}$. Thus,

$$-\log \sum_{s=t}^{N} p_{kc(i,s)} \quad \geq \quad -\log \sum_{s=t}^{N} p^{(l)}_{kc(i,s)} + 1 - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}}.$$

It follows that, up to a constant,

$$Q(p_{kj}) \quad \geq \quad q \quad = \quad \sum_{k=1}^{K} \sum_{i=1}^{M} \hat{z}_{ik} \log \pi_k + \sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \left(\log p_{kc(i,t)} - \frac{\sum_{s=t}^{N} p_{kc(i,s)}}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}} \right).$$

By iterative maximization of the surrogate function q we produce a sequence of p_{kj} (and of the mixing proportions π_k) values which have monotonically increasing Q value. The values converge to the maximum of Q with respect to p_{kj} and π_k.

Differentiation of q with respect to p_{kj} gives

$$\frac{\partial q}{\partial p_{kj}} \quad = \quad \sum_{i=1}^{M} \hat{z}_{ik} \left(\sum_{t=1}^{n_i} \frac{1}{p_{kc(i,t)}} \mathbf{1}_{\{j=c(i,t)\}} - \sum_{t=1}^{n_i} \frac{1}{\sum_{s=t}^{N} p^{(l)}_{kc(i,s)}} \mathbf{1}_{\{j \in \{c(i,t),\ldots,c(i,N)\}\}} \right).$$

We denote

$$\omega_{kj} = \sum_{i=1}^{M} \hat{z}_{ik} \sum_{t=1}^{n_i} \mathbf{1}_{\{j=c(i,t)\}}$$

and

$$\delta_{ijt} = \begin{cases} 1 & \text{if } j \in \{c(i,t), \ldots, c(i,N)\} \\ 0 & \text{otherwise} \end{cases} .$$

Therefore,

$$\frac{\omega_{kj}}{p_{kj}} = \sum_{i=1}^{M} \hat{z}_{ik} \sum_{t=1}^{n_i} \left[\sum_{s=t}^{N} p_{kc(i,s)}^{(l)} \right]^{-1} \delta_{ijt}$$

which implies that

$$p_{kj}^{(l+1)} = \frac{\omega_{kj}}{\sum_{i=1}^{M} \sum_{t=1}^{n_i} \hat{z}_{ik} \delta_{ijt} \left[\sum_{s=t}^{N} p_{kc(i,s)}^{(l)} \right]^{-1}},$$

for $k = 1, \ldots, K$ and $j = 1, \ldots, N$.

Similarly, maximization of q with respect to π_k and subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$ yields

$$\pi_k = \frac{\sum_{i=1}^{M} \hat{z}_{ik}}{M}$$

.

References

Aitkin, M., Anderson, D. & Hinde, J. (1981), 'Statistical modelling of data on teaching styles (with discussion)', *Journal of the Royal Statistical Society, Series A: General* **144**, 419–461.

Banfield, J. D. & Raftery, A. E. (1993), 'Model-based Gaussian and non-Gaussian clustering', *Biometrics* **49**(3), 803–821.

Bensmail, H., Celeux, G., Raftery, A. E. & Robert, C. P. (1997), 'Inference in model-based cluster analysis', *Statistics and Computing* **7**, 1–10.

Clancy, P. (1995), *College Entry In Focus: A Fourth National Survey of Access to Higher Education*, Higher Education Authority, Dublin, Ireland.

Critchlow, D. E. (1985), *Metric methods for analyzing partially ranked data*, Lecture Notes in Statistics, 34, Springer-Verlag, Berlin.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *J. Roy. Statist. Soc. Ser. B* **39**(1), 1–38. With discussion.

Diaconis, P. (1988), *Group representations in probability and statistics*, Institute of Mathematical Statistics, Hayward, CA.

Fraley, C. & Raftery, A. E. (1998), 'How many clusters? Which clustering method? - Answers via Model-Based Cluster Analysis', *Computer Journal* **41**, 578–588.

Fraley, C. & Raftery, A. E. (2002), 'Model-based clustering, discriminant analysis, and density estimation', *J. Amer. Statist. Assoc.* **97**(458), 611–612.

Hunter, D. R. (2004), 'MM algorithms for generalized Bradley-Terry models', *Ann. Statist.* **32**(1), 384–406.

Hunter, D. R. & Lange, K. (2004), 'A tutorial on MM algorithms', *Amer. Statist.* **58**(1), 30–37.

Hyland, A. (1999), *Commission on the Points System: Final Report and Recommendations*, Commission on the Points System Reports, The Stationery Office, Dublin, Ireland.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes factors', *Journal of the American Statistical Association* **90**, 773–795.

Keribin, C. (1998), 'Estimation consistante de l'ordre de modèles de mélange', *C. R. Acad. Sci. Paris Sér. I Math.* **326**(2), 243–248.

Keribin, C. (2000), 'Consistent estimation of the order of mixture models', *Sankhyā Ser. A* **62**(1), 49–66.

Lange, K., Hunter, D. R. & Yang, I. (2000), 'Optimization transfer using surrogate objective functions', *J. Comput. Graph. Statist.* **9**(1), 1–59. With discussion, and a rejoinder by Hunter and Lange.

Leroux, B. G. (1992), 'Consistent estimation of a mixing distribution', *Ann. Statist.* **20**(3), 1350–1360.

Lynch, K., Brannick, T., Clancy, P. & Drudy, S. (1999), *Points and Performance in Higher Education: A Study of the Predictive Validity of the Points System*, number 4 *in* 'Commission on the Points System Research Papers', The Stationery Office, Dublin, Ireland.

Marden, J. I. (1995), *Analyzing and modeling rank data*, Chapman & Hall, London.

McLachlan, G. J. & Peel, D. (2000), *Finite Mixture models*, John Wiley & Sons, New York.

Murphy, T. B. & Martin, D. (2003), 'Mixtures of distance-based models for ranking data', *Computational Statistics and Data Analysis* **41**(3–4), 645–655.

Plackett, R. L. (1975), 'The analysis of permutations', *Applied Statistics* **24**(2), 193–202.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Titterington, D. M., Smith, A. F. M. & Makov, U. E. (1985), *Statistical analysis of finite mixture distributions*, Wiley, Chichester.

Train, K. E. (2003), *Discrete Choice Methods with Simulation*, Cambridge University Press, Cambridge.

Tuohy, D. (1998), *Demand for Third-Level Places*, number 1 *in* 'Commission on the Points System Research Papers', The Stationery Office, Dublin, Ireland.