Group Discovery with Multiple-Choice Exams and Consumer Surveys: The Group-Question-Answer Model

Andrés Corrada-Emmanuel, Ian Beatty, and William Gerace
Computer Science Department and
Scientific Reasoning Research Institute
University of Massachusetts at Amherst
Amherst, MA 01003 USA
corrada@cs.umass.edu,{beatty,gerace}@physics.umass.edu

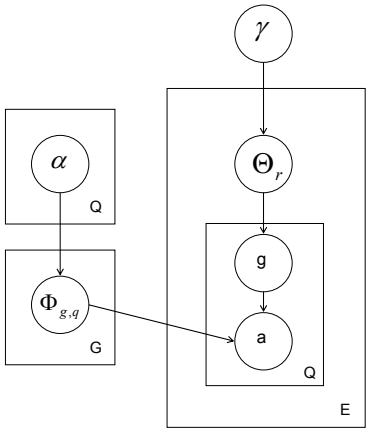
September 18, 2007

Abstract

Multiple choice questions (MCQs) are a common data gathering tool. We extend the Latent Dirichlet Allocation (LDA) framework to a collection of MCQ surveys. Topic discovery is turned into group discovery based on survey response patterns. Question choices are equivalent to vocabulary words and are conditioned on the question and the latent group that is used to cluster the survey responders. The structured format of MCQ surveys creates correlations between document 'authors' not found in unstructured natural language documents. We demonstrate the utility of the model by considering two performance measures : How well can we predict held-out question answers? What is the discriminatory power of the survey questions? The model should be of interest to anybody that uses MCQ surveys or exams to identify social groups.

1 Introduction

Multiple choice questions (MCQ) are a popular format for exams in large enrollment classes. The questions are easy to grade with automatic methods and are therefore sometimes the only economically viable format for testing a large number of students. The same economies of scale make the MCQ format a popular one for surveys that are interested in determining the tastes of a large group of consumers. The data collected during a survey or test could be viewed as a structured document collection where answers/words are correlated across documents by the question label. This motivates us to construct a generative model using the Latent Dirichlet Allocation (LDA) framework that has recently seen wide applicability to a variety of unstructured or semi-structured natural language document collections.



(a)

Symbol	Description
$\Theta_r \mid \gamma$	Respondent's group multinomial generated by Dirichlet(γ)
$\Phi_{g,q} \mid \alpha_q$	Group g answer multinomial for question q generated by Dirichlet(α_q)
$a_{q,e}$	Answer for question q on exam e
$g_{q,e}$	Group assignment for question q on exam e
G	Number of groups
E	Number of exams
Q	Number of questions
A	Number of answer choices

(b)

Figure 1: Graphical model representation of GQA and description of variables

The Latent Dirichlet Allocation (LDA) framework uses latent labels modeled as Dirichlet distributions to generate the multinomial response pattern of observed labels. The mathematical property of conjugacy between the Dirichlet and multinomial distributions has been known for some time, but it was not until the work of Blei, Ng, and Jordan [1] that the usefulness of this relationship was used to construct a generative model that clustered documents into latent 'topics'. The general idea of latent Dirichlet allocation has turned out to be incredibly useful in clustering many other document models [2, 3, 4, 5], and even images collections[6]. For that reason, we refer in this paper to the LDA framework as the generic term for using latent Dirichlet allocation.

The LDA model presented here has a novel feature. The exam or survey instrument is, in essence, a classifier or detector or groups. By modeling data collected with a specific exam, we can use the likelihood of a given questions answers conditioned on all other questions to rate the effectiveness of the questions themselves. This suggests that the LDA framework could be extended to sensors networks to provide autonomous assessment of detector reliability. We will also consider the task of predicting held-out question answers. This is possible with MCQ documents because the vocabulary is so small (typically 5 or 6 choices at most).

2 The Group-Question-Answer Model

The Group-Question-Answer model is a directed graphical model that clusters respondents to a survey based on the pattern of their responses to specific questions. It can accommodate survey instruments that have varying number of choices depending on the question. In this paper we detail the model for a MCQ survey with the same number of choices for each question. We illustrate the model in figure 1. Since GQA is a generative model we will describe it by considering the following pseudo-algorithm for constructing a synthetic dataset. We use the language of an exam to present the process.

1. Select the number of exams, and the number of questions in each exam.
2. Each question is modeled as a Dirichlet distribution that generates multinomials of student responses. Each question's Dirichlet distribution is parametrized by a vector $\vec{\alpha}$ of length equal to the number of choices for that question. Like all LDA models, this means the GQA model has no semantics built into it. For example, there is no need to know which is the correct choice for a specific question in the exam.
3. Select the number of latent groups.
4. For each group, generate the multinomial response pattern for each question by sampling from that question's Dirichlet distribution.
5. For each exam, draw from a Dirichlet distribution that governs group membership. The Dirichlet distribution is parametrized by $\vec{\gamma}$ of length equal to the number of groups.
6. For each question in the exam, sample the group membership multinomial for that exam and assign the group label to the question.
7. Given the group label for an exam question, draw a sample from that group's question distribution.

Putting all this together, the likelihood of producing a given exam collection is:

$$p(\{\Theta_r\},\{\Phi_{g,q}\},\mathbf{r}^e,\mathbf{q}^e,\mathbf{a}^e,\mathbf{g}^e \,|\, \boldsymbol{\gamma},\{\boldsymbol{\alpha}_q\}) = \left[\prod_{r=1}^{|R|} p(\Theta_r|\boldsymbol{\gamma})\right]\left[\prod_{g=1}^{|G|}\prod_{q=1}^{|Q|} p(\Phi_{g,q}\,|\,\boldsymbol{\alpha}_q)\right]\left[\prod_{e=1}^{|E|}\prod_{q=1}^{|Q|} p(g_{q,e}\,|\,r)p(a_{q,e}\,|\,\Phi_{g,q})\right] \quad (1)$$

2.1 Gibbs sampling update equation

Like other LDA models, equation 1 is not solvable in a closed form. There are a variety of methods for performing statistical inference with LDA models: the original variational approach in [1], collapsed variational [7], Gibbs sampling [8] and collapsed Gibbs sampling [7]. We used Gibbs sampling for the experimental results of this paper.

Gibbs sampling is based on updating latent labels (a group label in our model) by using conditional probabilities for labels on a given data point given labels on all other data points. The update equation for group assignment of each observed question answer is

$$P(g_i,a_i\,|\,\mathbf{g}_{-i},\mathbf{a}_{-i}) = \frac{(n^{(-i)}_{g_i,q_i,a_i}+\alpha_{q_i,a_i})(n^{(-i)}_{r_i,g_i}+\gamma_{g_i})}{\left(\left[\sum_{a'=1}^{A} n^{(-i)}_{g_i,q_i,a'}+\alpha_{q_i,a'}\right]\right)\left(\left[\sum_{g'=1}^{G} n^{(-i)}_{r_i,g'}+\gamma_{g'}\right]\right)} \quad (2)$$

The $n^{(-i)}$ notation is meant to represent that the counts are calculated excluding the student's question under consideration.

2.2 Finding the optimal α and γ

The full GQA model has one Dirichlet parameter vector, $\boldsymbol{\gamma}$, which is G long and Q Dirichlet parameters, $\{\boldsymbol{\alpha}_q\}$, each A long. To simplify our experiments, we took the canonical choice of uniform Dirichlet parameters: $\boldsymbol{\gamma} = \gamma\vec{1}$ and $\boldsymbol{\alpha}_q = \alpha\vec{1}$.

Equation 1 can be rewritten in terms of the counters $n_{r,g}$ and $n_{g,q,a}$ that are used in the update equation 2. In the case of uniform Dirichlet parameters, the rewritten equation is,

$$P(\mathbf{r}, \mathbf{q}, \mathbf{a}, \mathbf{g} \,|\, \boldsymbol{\gamma}, \boldsymbol{\alpha}) =$$
$$R\left(\ln\Gamma(G\gamma) - G\ln\Gamma(\gamma)\right) + \sum_{r=1}^{R}\left(\sum_{g=1}^{G}\ln\Gamma(n_{r,g}+\gamma) - \ln\Gamma(\sum_{g=1}^{G} n_{r,g} + \gamma)\right) +$$
$$GQ\left(\ln\Gamma(A\alpha) - A\ln\Gamma(\alpha)\right) + \sum_{g=1}^{G}\sum_{q=1}^{Q}\left(\sum_{a=1}^{A}\ln\Gamma(n_{g,q,a}+\alpha) - \ln\Gamma(\sum_{a=1}^{A} n_{g,q,a} + \alpha)\right) \quad (3)$$

This equation can then be used to perform a gradient ascent search for the parameter settings that maximize the log likelihood of the observed data.

2.3 Answer prediction

The small number of choices commonly found in MCQ surveys makes it possible to measure the answer prediction performance of the GQA model. The LDA framework has been applied to documents with a large vocabulary, more than a thousand words being typical. With such a large vocabulary, it would be difficult to detect any improvements in word prediction. A word is either correctly predicted or it is not. An LDA model prediction would then be hardly better than a uniformly random predictor on documents with large vocabularies.

Since survey documents have a vocabulary of six or so words, it becomes practical to use the GQA model to predict held-out answers. This prediction can then be compared to various other predictors. The equation for GQA answer prediction is constructed easily from the posteriors obtained from MCMC runs as

$$P(a \,|\, r, q) = \sum_{g=1}^{G} P(a \,|\, g, q) P(g \,|\, r) \quad (4)$$

This prediction can be compared to three other predictors: the uniformly random guesser, the average respondent choice guesser, and the average question choice guesser.

2.4 Assessing the quality of the questions

A perennial question in survey and exam making is the problem of assessing the quality of the questions used to carry out the survey. This problem is related to the question of detector self-assessment. Within the context of exams one can say that there are two failure modes for questions. A question may be so hard that student response is essentially uniformly random or so easy that all students have the same correct response.

One way to quantify the quality of exam questions is to rank them according to their likelihood given all other questions, specifically the quantity $P(\mathbf{a}_q, \mathbf{g}_q \mid \mathbf{a}_{-q}, \mathbf{g}_{-q})$. The same mathematics that results in the Gibbs sampler update equation can be used to calculate this likelihood as,

$$\left[\prod_{g=1}^{G} \frac{\Gamma(\sum_{a'=1}^{A} \alpha_{q,a'})}{\prod_{a'=1}^{A} \Gamma(\alpha_{q,a'})} \frac{\prod_{a'=1}^{A} \Gamma(n_{g,q,a'} + \alpha_{q,a'})}{\Gamma(\sum_{a'=1}^{A} (n_{g,q,a'} + \alpha_{q,a'}))}\right] \times \prod_{r=1}^{R} \frac{n_{r,g=g_{r,q}} + \gamma_{g=g_{r,q}} - 1}{\sum_{g'}^{G} (n_{r,g'} + \gamma_{g'}) - 1} \quad (5)$$

Questions can then be ranked according to this likelihood. The question with the highest likelihood given all other question responses being the least discriminating one.

3 Experimental Results

We applied the GQA model to an exam given to a large science class. The exam consisted of twenty questions with 6 choices per question[1] and was taken by 230 students. Dirichlet parameter selection was optimized by maximizing the likelihood on ten one-held-out-question datasets. A one-held-out dataset was created by randomly picking one question from a student's exam to be used for testing. The student's remaining questions constituted the training set. Ten realizations of this random procedure created the datasets used for the Dirichlet parameter searches. A Gibbs sampler using random scanning was run for a small number of groups ($G = 2, 3, 4, 5$).

3.1 Answer Prediction Results

We carried out answer prediction for the $G = 2, 3, 4, 5$ maximum likelihood models trained on the one-held-out datasets. The LDA framework does not have any 'semantics' built into it. In the case of GQA, that means that it is not necessary to know what the correct choice is for any of the exam questions. This is why the model is equally applicable to consumer surveys where there is no correct choice. But since students are trying to maximize their exam score, there is a bias toward certain answers. The exam dataset we use in this paper has the property that eighteen of the questions have 'A' as their correct choice (see section 5 for why this is the case). This suggests two very good, i.e. much better than random, answer predictors for this dataset.

We call the first predictor *average student choice*. For all the training questions answered by the student, a multinomial of the answer choices is constructed. The second predictor is *average question choice*. A multinomial for a given question is constructed across all student responses in the training set. The comparison between the GQA predictor, average question choice, and average student choice are shown in table 1. The performance was measured by

[1]There were 5 choices for each exam question but we have added a sixth choice to represent the questions that received no answer on the assumption that not being able to answer a question during an exam is informative about group labels. This design choice was confirmed by the results in section 3.2

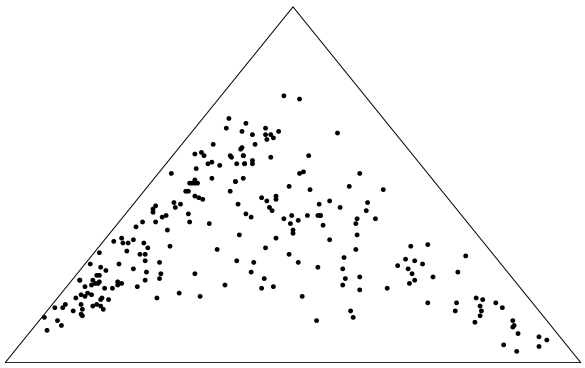


Figure 2: Students group posteriors for the maximum likelihood solution of GQA for $G = 3$.

Number of groups	GQA	Average question choice	Average student choice
2	56.88	53.20	44.02
3	56.41	53.26	44.21
4	55.40	53.24	44.14
5	53.47	53.51	44.14

Table 1: Percentage of held-out answers correctly predicted.

averaging over 40 sampling runs over the 10 random one-held-out datasets since there is no way to ascribe a distance between predicted versus observed answers. [2] The average question and student choice predictors should be constant across varying number of groups and their values were used across groups to determine the precision of the answer predictors (in this case about $0.2 - 0.3\%$) given our finite sampling.

These results for answer prediction show that the exam data seems to be explainable by at most 2 or 3 groups. We examined the student clusters for $G = 3$ as shown in Figure 2. The cluster on the left corner consists of students that answered most questions correctly. The cluster on the right corner consists of students that got most questions wrong, and the upper corner consists of students that got most questions right but failed on some questions. LDA models are typically used to create 'topic' lists. While suggestive, these lists are hard to quantify or compare with the results between different LDA variants such as Author-Topic [2, 3] or Author-Recipient-Topic [4]. These results are a further demonstration that 'topics' or 'groups' have an empirical meaning given a particular dataset. The question of how many latent topics or groups are optimal for a given dataset is an unsolved problem in the LDA framework. Like other clustering algorithms, the LDA framework has no mechanism for determining the optimal number of clusters given a specific dataset.

3.2 Question Assessment

Question assessment for this exam dataset was carried out by considering all of the exam data. We used a $G = 3$ fitted model with $\alpha = 0.99$ and $\gamma = 1.51$. Some selected questions ranked

[2] This is not the case for questions that ask for a rating choice such as the Netflix Prize dataset.

Question Id	12	1	17	15	20
Likelihood	1st	20th	14th	15th	11th
% correct	1st	19th	20th	7th	18th

Table 2: Questions ranked by likelihood and percentage of correct responses

by their likelihood and percentage of correct responses are shown in Table 2. Some of the rankings are as we expected but others were not. The least informative question for GQA was question 12. It was also the question that received the most correct responses. This question corresponds to one of the extreme failure modes for an exam question – it was so easy that 217 out of the 230 students answered it correctly.

But the most informative question for the GQA model, question 1, was not the hardest question in the exam. It received 99 correct responses while question 17 received 64 correct responses. The reason question 17 was not as informative is that many students gave the same wrong answer – choice 'B'. The best performing latent group had a posterior probability of 30% of picking it, the worst performing group, 33% and the middling group, 86%. In fact, choice 'B' was the most common choice for this question. This is similar to the case of question 12. Picking the same answer, for whatever reason, makes a question uninformative.

The most extreme difference in the two methods of ranking the questions occurred for question 15. It was ranked by GQA eight spots above what one would expect by the its difficulty. It received 176 correct responses. While the best performing group and the middling group had high probability of answering it correctly (92% and 84%), it was the question that was most unanswered. The worst performing group had a posterior probability of 5% of not answering it, the middling group, 2%, and the best performing group, 0.5%.

These question assessment results are typical of what is observed with other LDA models. Their application leads to intuitive results such as the least informative question being the one most likely to be answered correctly. But they also uncover patterns or rankings that one would not expect but which are reasonable upon further investigation, such as the case of question 15 discussed on the previous paragraph.

4 Related Work

The use of the LDA framework to discover social entities was first considered within the context of a collection of emails from the Enron corporation [4]. Explicit group discovery was also considered by Wang et al [5] within the context of voting blocs in the United Nations Assembly and the United States Senate. The paper also has an informative comparison of group discovery with the LDA framework versus Blockstructure models. We can summarize the difference as follows. Blockstructure models find groups by considering the linkage between entities. In the context of a survey or exam, the stochastic blockstructure model in [9] could be applied with the value of a linkage equal to the number of questions the respondents answered similarly. LDA models such as Group-Topic [5] and GQA in this paper perform group discovery by considering the attributes of the linkages. Knowing what answers respondents shared is more informative than just knowing they answered n questions similarly.

The similarities and differences between the GT model [5] and GQA are also worth con-

sidering. Voting for bills is similar to picking an answer to a question with two choices. Respondents to surveys, however, do not 'vote' to oppose or agree with other respondents. Thus, a large component of exam taking or survey filling is not political but consumer oriented. Consumers are answering according to their taste with little or no social interaction during the survey. Consumers certainly have preferences that are conditioned on other social attributes (education, class, income, etc.) but their survey responses should be weakly coupled to strategic voting in agreement or opposition to other social groups. GQA neglects any such social interactions.

5 An Exam Cheating Model

The notable exception to this consumer viewpoint for exams occurs when students do interact during an exam, i.e. they cheat. Assuming perfect, non-collusive cheating we can create a cheating exam model. Each question is assigned a Beta distribution parametrized by κ_q that generates a binomial for cheating for each student, $K_{r,q}$. The likelihood of the observed exam responses is then given by

$$p(\{\Theta_r\},\{K_r\},\{\Phi_{g,q}\},\mathbf{r}^e,\mathbf{q}^e,\mathbf{a}^e,\mathbf{c}^e,\mathbf{g}^e \,|\, \boldsymbol{\gamma},\{\boldsymbol{\alpha}_q\},\{\kappa_q\}) = \left[\prod_{r=1}^{|R|} p(\Theta_r \,|\, \boldsymbol{\gamma})\right]\left[\prod_{r=1}^{|R|}\prod_{q=1}^{|Q|} p(K_r \,|\, \kappa_q)\right]\left[\prod_{g=1}^{|G|}\prod_{q=1}^{|Q|} p(\Phi_{g,q} \,|\, \boldsymbol{\alpha}_q)\right] \left[\prod_{e=1}^{|E|}\left(\prod_{\text{honest}} p(g_{q,e} \,|\, r)p(a_{q,e} \,|\, \Phi_{g,q}) \prod_{\text{cheating}} p(c_q \,|\, K_{q,r})\right)\right], \quad (6)$$

where *cheating* refers to the subset of student questions latently labeled as 'cheats', conversely for *honest*. The perfect part of the cheating comes from the cheater putting down the correct answer when they cheat. The non-collusive part of the cheating is expressed by conditioning the probability of cheating solely on the question and the respondent, neglecting any group interaction.

A common procedure for minimizing cheating during MCQ exams is to produce randomized question and choice order versions of the same exam. Question 10 in one exam is question 3 in another one. The correct choice for a question in one version is 'a', in another version, 'c'. The exam data previously discussed was given under precisely this randomized protocol.

This suggests the following protocol to statistically measure non-collusive, perfect cheating during an exam. A large class could be divided into two random groups. One group is given the randomized versions of the exam, the other is given a single version of the exam. The GQA model could be fitted to the random-order exam group by maximizing the likelihood of the observed responses. These GQA Dirichlet parameters are then frozen and the $\{\kappa_q\}$ are varied to maximize results on the homogeneous exam group. The resulting posterior probabilities for cheating can then be used as an estimate for the rate of question cheating.

6 Conclusions

We have presented a Latent Dirichlet Allocation model for group discovery with multiple choice exams or consumer surveys. The practical utility of the model was shown by analysing an exam given to a large class (230 students). The percentage of correctly predicted held-out question answers with the GQA model outperforms other possible predictors. In addition, we considered a novel task for LDA models: using the results of the survey to assess the group discriminatory power of the questions themselves. The effectiveness of some questions was roughly correlated with their degree of difficulty as measured by the number of students that answered it correctly. But some question rankings were surprising, one question being relatively easy yet still effective in differentiating between the latent groups being discovered. This is an approach that could be extended to perform autonomous self-assessment of a collection of detectors. In addition, we introduced an exam cheating model and protocol to statistically measure question cheating rates *ex post facto*.

References

[1] David M. Blei, Andrew Y. Ng, and Michael J. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. Griffiths. Probabilistic author-topic models for information discovery. In *The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004.

[3] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 2004.

[4] Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 786–792, August 2005.

[5] Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and their attributes. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*. MIT Press, Cambridge, MA, 2007.

[6] Josef Sivic, Bryan C. Russell, Alexei A. Efros, Andrew Zisserman, and William T. Freeman. Discovering object categories in image collections. Technical report, MIT Computer Science and Artificial Intelligence Laboratory, 2005.

[7] Yee Whye Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press, Cambridge, MA, 2007.

[8] Thomas L. Griffiths and Mark Steyvers. Finding scientific topis. In *Proceedings of the National Academy of Sciences*, volume 101 (suppl. 1), pages 5228–5235, 2004.

[9] Krzysztof Nowicki and Tom A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087, 2001.