Bayesian Item Response Theory and Latent Dirichlet Allocation applied to textual professor performance profiling

by

Eduardo Jorquera

Advisor: **Harvey Rosas**, Ph.D.

Co-advisor: **Álvaro Montenegro**, Ph.D.

A thesis submitted to the faculty of the Statistics Institute

in partial fulfillment of the requirements for the degree of

Master in Statistics

Institute of Statistics

School of Natural Sciences

Universidad de Valparaíso

2018

Jorquera, Eduardo (M.S., Statistics)

Bayesian Item Response Theory and Latent Dirichlet Allocation applied to textual professor performance profiling.

Thesis directed by Harvey Rosas, Ph.D. and Álvaro Montenegro, Ph.D.

The evaluation of teaching and the perception about the academic environment are only obtained through students. Measure the teachers' performance and subject relevance has an increasing importance for universities and academic institutions since these evaluations are subjective. Thus, a natural question is: how to develope an unbiased measure through subjective answers? This thesis uses several statistical tools in order to measure professor performance and subject relevance. An strategic diagram is constructed to compare these latent variables. Item response theory (IRT) is used to propose a method to measure teachers' performance and subject relevance, also Latent Dirichlet allocation (LDA) is used to model textual data, and clustered in order to profile professors using students' comments.

keywords: evaluation of teaching, strategic diagram, Item response theory, Latent Dirichlet allocation, latent variables.

To my family

Acknowledgements

Thanks to my family and friends, specially to professor Harvey, who always advised me as a friend. Thanks to each professor who helped me and contributed in my academic formation.

Specially, thanks to God, who held me and showed me the path I must follow.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

There are a lot of ways to think about how to evaluate "how good" or "how bad" a teacher and student evaluation of teaching (SET), this has been a subject of constant debate for a long time (Stark and Freishtat, 2014). Also, SET is often used to make decisions related to professors careers; furthermore, the only feedback from students are these evaluations. Even when studies repeatedly display that SET ratings are currently explaining less than 50% of the variability of student learning, and the correlation between these measures and other good teaching metrics are not high enough to not to improve SETs (Becker and Watts, 1999).

Uttl et al. (2017) found some interesting results about SETs: students who perceive themselves with a good workload and with enough amount of time for learning, will give higher SETs. Given that students who feel confident about their knowledge for an specific subject, will evaluate differently than students who do not have that confidence. On the contrary, those professors who lack confidence in their own skills either not in conditions to do it properly or they decide to not do it because he believes that such students in particular are detrimental to themselves and/or for the society in long term, then the professor will get a low SET score. Other aspect in their findings, is the fact that professors with high SETs scores, does not imply their students will necessary learn more. Furthermore, the correlation between learning and SET scores is equivalent to generate correlations from a population with null correlation ($\rho = 0$) between both variables. Thus, there is not evidence to support that students learn more from professors with higher SETs.

Another interesting comparison is the expected grades of students and SET scores; Krautmann and Sander (1999) found that lowering grading standards will produce higher SET scores, which can be seen as "buying" a good SET score. Also, there is a significant relationship between the student's cognitive style and student's perceptions of teaching (Aigner and Thum, 1986), the same study showed that the tendency is to rate too generously, this affect an "impartial" analysis and SET scores are biased.

Dilts (1980) proposed a linear model to simulate evaluation averages, but this was improved by Mehdizadeh (1990), who introduced loglinear models. Exploring the use of this technique in categorical economic education variables. The author highlights the advantages to allow variety of interaction between variables. A logit link function is implied by a loglinear model, because logit models are a special case of loglinear models, where dependence between regressors variables is identified as well as the interaction of the dependent variables. After fitting a hierarchical model to the SET data, the expected grade, supplementary material, and availability of professors out of class schedule were found as important factors in SET scores.

It is important to understand the environment where the student develops; improving in this characterization, Seiver (1983) shows that a better faculty improves the performance of students grades; when SET programs are instituted, professors attempt to have better performance among their classes, feeling the pressure. There is an institutional factor related to SET scores, which means that to make a good model implies to deal with certain distortion control in evaluation process (Dilts, 1980).

Other interesting feature to study in the students' environment is the class size; related to this topic, Decanio (1986) found that class size is a negative influential variable over SET scores, mainly, in the middle range responses, these are categories b-"good" and c-"average", where a is

excellent and e is very poor. Other finding in this study, is the conclusion in agreement with Seiver (1983), that there is no significant relation between expected grades and SET scores.

Despite the criticism, SET is used by a huge number of universities to measure the evaluation of teaching effectiveness (Wachtel, 1998; Uttl et al., 2017). Commonly, SET is made at the end of the academic period, within the last two weeks of courses. When students answer the questionnaire about the perception of instructors and courses, often the instrument consists in a 5-point Likert scale.

Some assumptions to validate SET scores are questionnaires are anonymous, only students can observe and perceive professors' performance, and the questionnaire is truthfully answered. Stark and Freishtat (2014) suggest that SET scores are influenced by students' characteristics, such as gender, ethnicity, how attractive they find their professor. But, this may be affected because the students' objectives may differ from those coming from faculty and management. Students can also evaluate differently a teacher who in charge of a core course for their program than a teacher in charge of a complementary one, this could be a problem since teachers of advance courses may be seen as more capable than introductory ones (Dilts, 1980).

To analyze SET scores data, this thesis propose to use Rasch models to analyze categorical data, and Latent Dirichlet allocation for students' comments related to professors' performance.

1.1 Hypothesis

The actual teaching and subject evaluation can be improved though applying IRT model to perform professors and subject measures and LDA to summarize students' comments.

1.2 Objectives

1.2.1 General objective

Measure professors' performance and subject relevance through students' evaluation.

1.2.2 Specifics objectives

(1) Measure students' evaluation for teachers.

(2) Measure subjects' relevance among the undergraduate programs.

(3) Summarize students' comments about professors.

To develop the intuition behind IRT, Rasch models will be described; first, the dichotomous case and then the multi-faceted Rasch model are described in Chapter 2. Also in Chapter 2, unidimensionality analysis, Kendall correlation, mutual information, and goodness of fit are described in order to be used later. LDA is described in Chapter 3, where estimation and theoretical background for text analysis is theoretically developed. The chapter includes the intuition behind the Dirichlet allocation model. Also, Gibbs sampling is described and developed very intuitively. Chapter 4 describes how the data was gathered. Also, the steps to analize the categorical and textual data analysis are included. In Chapter 5, the results from analysis are displayed. Finally, conclusions and discussions can be found in Chapter 6, and discussing main results.

Chapter 2

Item response theory for SET data

Tests have a wide range of applications among schools, industry, psychology, and others. Several students' factors can determine SET scores. In general, tests may have a variety of functions, and a common use of these is to classify and measure.

2.1 Rasch models and IRT

In different fields, it is of the utmost importance to apply test regularly; these have been used for a lot purposes, such as educational systems (SET case), learning process, ability measurement, intelligence, between others. Awareness of its importance, limitations and impact of testing has implied to criticism. These facts were useful for the introduction of new methods to analyze tests scores.

2.1.1 Item response data structures and hierarchically structured data

The main difference of this kind of data structure, is that the data is obtained in obtained by respondents. This in a statistical point of view, is such a characteristic since other commonly used of capturing data is made by a "data capturer"; this person may be a doctor, a chemist, etc.

Another main feature of IRT is that difference between respondents can be modeled by a

probability distributions; and inference can be done since the parameter's population distribution, which will be decisive for this thesis.

The standard situation is to assume a sample where every responder responses are independent from each other, as well as sampled from an infinite population with replacement. Commonly, responders belong to subpopulations given certain clustering, where observations are correlated within clusters. It is said that observations are hierarchically structured when nested in clusters. When observations belong to the same cluster, typically they are not independently distributed. On the contrary case, observations from different clusters (non nested) are typically independently distributed.

IRT is not the only case when it is about clustered (response) data. A common case is longitudinal data, when subjects are measured repeatedly on the same outcome for several points in time. For longitudinal data, when the number of measurements and spacing of time vary between the subjects, then the observations are viewed as nested with subjects. Another known term is repeated measurements, but it refers to data on subjects measured repeatedly in different conditions and time. Hierarchical structured nature of the data is characterized by clusters, which can be nested in characteristics of interest from the population. This kind of nested characteristic may be for example geographical, political, belongingness to certain subject class, etc. For educational cases, the response data is often nested to more than one characteristic, commonly, nested within individuals and are in turn nested within organizations (for instance school, faculty, university). For multivariate data case, data has a hierarchical structure, since for each subject there are multiple outcomes which are measured; thus is nested within the subject.

For hierarchically structured data there are different terms, for each level of hierarchy; level-1, stage-1, micro or observational level means the lowest level of the hierarchy. For higher levels of hierarchy, level-2, stage-2, macro or cluster level terms are used. The appropriate terminology will

depend of the context, and when there is an absence of the context, the most generic terms are level-1 and level-2. This concept can be explained in the following way, the level-2 of the hierarchy is the nested groups of respondents, that is, the "big cluster" where the respondents belong to. This may be the school, country, etc. On the other hand, level-1 cluster is nested within the individuals. Therefore, inference has to be made at different levels of aggregation, through a statistical model which has to comprise different levels hierarchically. Conditional likelihood is used to model a within-respondent level, where conditional independence is typically assumed given a person parameter. There is heterogeneity between respondents for higher levels of the hierarchy. In order to analyze the hierarchically structured response data a bayesian modelling approach is developed. A bayesian modelling approach will be developed for hierarchically structured response data, in order to analyze it.

Often response data is sparse at the respondent level, and linked to many respondents. The sparsity complicates the estimation procedure of obtaining reliable estimates for individual effects. Better estimation can be obtained from individual effects taking into account the response data from other individuals nested in the same group. In the same way, more accurate estimations at an aggregate level by using within individual data.

Commonly response data are made by integer values, where responses can be considered as correct or incorrect, or obtained using a 5-7 point scale. The lumpy nature of response data requires special model approach, since standard distributional assumptions do not apply for this case.

As typically as it is, response data is obtained in combination with other input variables. For example, response data is obtained from joining the collected responses and grading records. Here the objective is to make a joint inference over individual and scholar effects given a resulting output variable. With a bayesian background, different information sources can be managed in a efficient way, considering certain uncertainty. It will be showed that a bayesian approach together

with a powerful computational resources offers a good alternative to analyze response data.

2.1.2 Latent variables

Multiple definitions can be found about latent variables in the literature. In this section, latent variables will be defined as a random variables, where their realizations are not directly measured, that is, there is no directly way to observe them. A direct implication of this, is that the latent variable can not be directly measured, as is the case of measuring intelligence, ability or motivation. An operative definition establishes that its construction is related to observable data. Often the relationship is defined such that item responses turn as indicators for underlying construct measurement. For example, common response models define a mathematical relationship among responses of a person and a latent variable which represents the person property that it measures. Even though questionnaires have a integer input (for example a Likert scale), the latent variable will be commonly a continue random variable. It is also possible that a latent variable be defined as categorical, which may be ordered or unordered. Bartholomew and Knott (1999) and Rabe-Hesketh and Skrondal (2004), among other authors, offer a general description of classical models for latent variables and its different applications for social sciences.

For a variety of reasons, latent variables play a important role in statistical models for response data. Specially in a behavioral and social investigation, given the nature of the research. First, as was mentioned before, is commonly assumed that item responses are indicators of an underlying construction or latent variable, and the main focus on its measurement. IRT defines a relationship between the element responses and the respondents latent variables. Second, the direct specification of a joint distribution of random observations is often hard. Latent variables are useful to reduce dimensionality through the definition of an underlying structure, and will be enough to explain relationships with a reduced variable set. In third place, is often the case that a partial observation is marked as an underlying continuous variable. For example, a common assumption

is that continuous latent variables are not observed, but discrete response data results can be obtained using a censoring mechanism. For binary responses, a positive response is captured when an underlying continuous latent variable is higher than a threshold value, and a negative response can be deduced for contrary case. The formulation of a continuous variable response is very flexible, as it will be showed in the rest of the chapter.

2.1.3 Traditional item response models

IRT cares about measurement of a hypothetical construction that is latent, and it only can be measured through another tractable variables. This hypothetical construction is a latent variable, and commonly represents ability, intelligence, or in general terms, a latent characteristic from the respondent. Within this document, the notation for this latent characteristic will be θ. When latent variable refers to a person characteristic, as an ability, also is known as person parameter.

Item response models have desirable features, most of these features come from the fact that a common scale is defined for latent variable. Questions and respondents features are both separately parameterized within an item response model and both are invariant. This means that corresponding estimation does not depend of the test. Latent variable estimations from different elements set whose measured at the same underlying construction are comparables and differ only at measurement error. Item characteristic estimation from different individuals responses from the same population are comparable and differ only in sampling error.

There are two key assumptions involved in IRT: the first assumption is that a change of latent variable will lead to a probability change from a response, and this is completely described through the item characteristic curve (ICC), the item characteristic function or the trace line. The ICC specifies how the item response probability changes given other latent variable changes. Different mathematical ways of characteristic curves from item lead to different item response models. For

dichotomous responses (correct or incorrect), the correct answer probability is modeled as a person and item parameters function. The second assumption states that a response to a pair of items are statistically independents when an underlying latent variable is constant (the elements measure an unidimensional latent variable). In this case, only an unidimensional latent variable influences the item responses and traditional items response models. When this assumption is true, there exists independence. The local independence assumption is generalized easily to a multidimensional latent variable, which indicates that responses to a pair of items are statistically independent when a multidimensional latent variable keeps as constant.

A random vector of K responses is denoted as $\mathbf{Y}_i$, with observed values $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iK})$ of a respondent indexed as $i-$th, with ability parameter θ_i. Then, the assumption of local independence can be formulated or written as

$$\begin{aligned} P(\mathbf{y}_i|\theta_i) &= P(y_{i1}|\theta_i)P(y_{i2}|\theta_i)\ldots P(y_{iK}|\theta_i) \\ &= \prod_{k=1}^{K} P(y_{ik}|\theta_i) \quad . \end{aligned} \tag{2.1}$$

Local independence assumption is also known as conditional independence, because there is one underlying variable when local independence exists. Thus, the term local independence is also known as conditional independence.

The equation (2.1) can be interpreted as the probability that the $i-$th respondent with ability θ_i produces $\mathbf{y}_i$ response (Holland, 1990; Molenaar, 1995). From an stochastic point of view, respondents have an stochastic nature, which means that to say that a respondent with ability θ_i has certain probability to produce a correct answer is meaningful. The idea of this is that each person gives little response variations when the respondent is confronted with the same item many times and produce a bias inside the respondent after each confrontation. Lord and Novick (2008) defined a propensity distribution $F_{ik}(y_{ik})$ of the random variable Y_{ik}, which is the response of the $i-$th individual at the $k-$th item. The existence of this distribution is hypothetical because is not

always possible to obtain a big number of respondents. Moreover it defines the true score of a person τ_{ik} as the expected value of the observed score:

$$\tau_{ik} \equiv E(Y_{ik}) \quad .$$

Holland (1990) remarked that from an stochastic point of view, the subject can suggest the lack of need of a model for a population model for respondents (examinee population), but the population effect always will be present (person and item parameters always will be estimated in a nested the population). This leads to reformulate the sampling concept of respondents from a population. In a random sample view, Rasch probability in the product at (2.1) is a proportion of respondents with ability θ_i whose response is correct. This point of view makes the respondents population as part of the probability model for every answer. Simple random sampling approach is adopted for conditional probability interpretation.

2.1.4 Binary item response theory

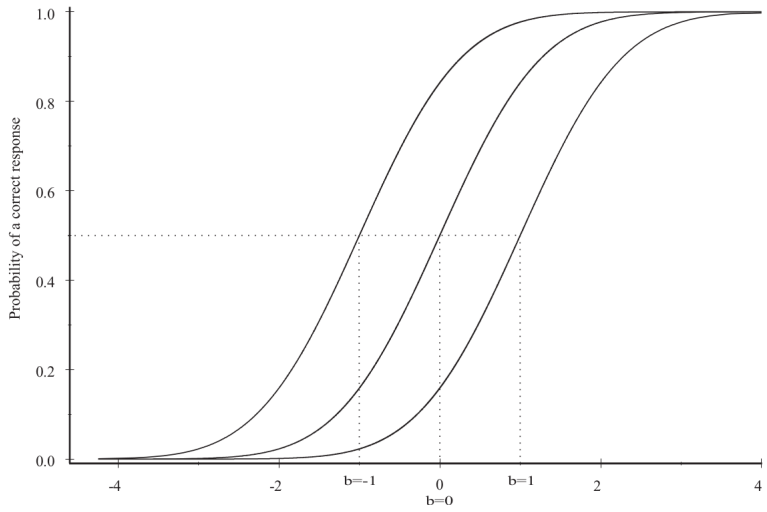

Figure 2.1: ICC for three curves related to three different difficulty levels, from (Fox, 2010).

Rasch models where introduced by Rasch (1960), who formulated a one-parameter logistic response model, and is one of the simplest and most widely used for response models. The probability of a correct response will be given by

$$P(Y_{ik}|\theta_i, b_k) = \frac{e^{\theta_i - b_k}}{1 + e^{b_k - \theta_i}} = \left(1 + e^{b_k - \theta_i}\right)^{-1} \quad , \tag{2.2}$$

where the ability θ_i of the $i-$th respondent, with item difficult parameter b_k. The plot of this equation can be seen in Figure 2.1. Thus, each ICC describes the relationship already existing between the ability and the probability of a correct answer. In the ability scale from the ICC, the difficulty parameter b_k indicates the point of the ability scale when correct response probability is 0.5. An interpretation of ICC is that if probability of success is higher compared to another item given the same level ability θ_i, it is said that the item is easier. In Figure 2.1, from left to right, the curves are increasing in difficulty. It can be easily seen that to keep 0.5 as the probability of a correct answer for each item, the ability parameter θ_i has to increase from -1 to 1. An important characteristic for ICCs for different items is that they are parallel to each other, since they are from a Rasch model. In other words, for this three elements, an increasing ability leads to the same probability of success. Furthermore, elements discriminate similarly between success probabilities for related ability levels. The discrimination of items is in the same way that probabilities of success discriminate for related ability levels.

Rasch (1960) developed in his models the description of how the dependent variable can be expressed as the log odds, or logit of passing an item; this is equal to the ability parameter minus the item difficulty, that is: $\theta_i - b_k$. As introduced before, Rasch models have desirable features. The probability distribution belongs to exponential family distributions, this implies that the distribution has mathematical and statistical properties of exponential family, such as consistency, completeness and minimal sufficiency, among others (Lehmann and Casella, 2006). As equation (2.2) shows, the Rasch model allows to separate the ability parameter θ_i from the difficulty parameter b_k. In fact, in parameter estimation, for ability parameter can be eliminated through using

conditional maximum likelihood (CML) estimation. This can be made when the response space is correctly partitioned according to the in addition of raw scores; for ability parameter θ_i is a sufficient statistic. Similarly, for item difficulty b_k, the item scores are sufficient statistics.

Through adding a constant to the ability parameter θ_i or subtracting this constant to the difficulty parameter b_k, it can be seen that for increasing success response probability (equation 2.2). Both parameters are defined in the same metric space, and the metric is defined up to linear shift. This problem can be solved specifying the restriction such as that metric location is known. This is generally done through equating the sum of difficulty parameters b_k is equal to zero, or restricting the scale to zero.

An assumption of Rasch models is that all items discriminate between respondents in the same way, since that items only differ between themselves in item difficulty. This is a limitation for the Rasch models, because from a practical approach, it is desirable to parameterize item difficulties and item discrimination. Looking up for the history of the model, Thissen (1982) introduced an estimation method through the Bock-Aitkin algorithm (Bock and Aitkin, 1981), which uses the expectation maximization (EM) algorithm. As EM algorithm uses marginal maximum likelihood (MML) to estimate one parameter logistic model, where all discrimination parameters are equal but they do not have the restriction to be equal to one.

2.1.4.1 Two-parameter model

The two parameter logistic model, has an item discrimination parameter a_k, which describes the slope or rate of change of the ability parameter θ_i, that is,

$$P(Y_{ik}|\theta_i, a_k, b_k) = \frac{e^{a_k\theta_i - b_k}}{1 + e^{a_k\theta_i - b_k}} = \left(1 + e^{b_k - a_k\theta_i}\right)^{-1} \quad . \tag{2.3}$$

The slope parameter a_k will affect the ICC, thus, ICC will not be necessarily parallel for different difficult parameters b_k. In Figure 2.2 three ICCs are plotted from two-parameter IRT model with the same difficulty ($b_k = 0$). Then, the discrimination parameter a_k characterize every ICC.

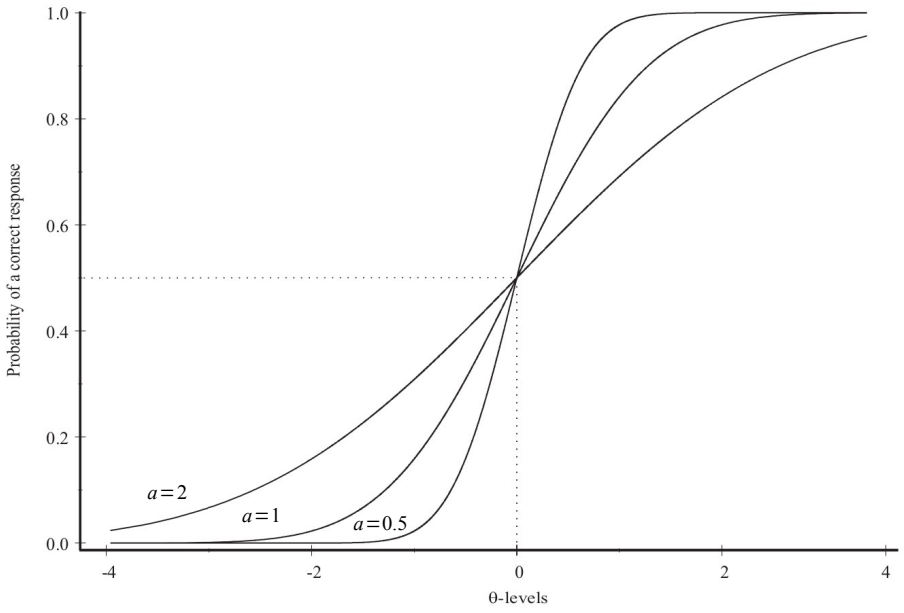

Figure 2.2: ICC for three curves related to three different discrimination levels, from (Fox, 2010).

The discrimination parameters in Figure 2.2 are: 2, 1 and 0.5. The steepest slope is when $a_k = 2$. The higher as the discrimination parameter is, the better the item is able to discriminate, and the same for lower discrimination parameters. Note that item discrimination parameter a_k has a close relation with item difficulty parameter b_k. A high discrimination item is useful only for the area of the item difficulty level, from the ability scale. This means that in Figure 2.2 the steepest slope item is useful in the region demarcated by -1 and 1, for ability scale values. Now looking at the item with the flattest ICC, the region between -2 and 2 will be the useful one.

For this model does not exists a sufficient statistic for ability parameters, and is not possible to make an estimator through maximum likelihood method. As was mentioned before, Bock and Aitkin (1981) proposed an algorithm through MML for the two parameter model. From the marginal distribution, integrating over ability distribution, which allows to remove the ability parameters from likelihood function.

The normal ogive model is defined in literature as a probit version of the two-parameter model (Lord and Novick, 2013), where the normal distribution can describe the ICC:

$$\begin{aligned} P(Y_{ik}=1|\theta_i,a_k,b_k) &= \Phi(a_k\theta_i-b_k) \\ &= \int_{-infty}^{a_k\theta_i-b_k}\phi(z)dz \quad , \end{aligned} \tag{2.4}$$

where $\Phi(\cdot)$ is the normal distribution functions and $\phi(\cdot)$ is the normal density function. Note that equations (2.3) and (2.4) are close when the logistic item parameter values are multiplied by a scale constant $d = 1.7$. Then, for different θ_i values, both equations differ at most in 0.01 in absolute value (Hambleton, 1991). Thus, the model can be rewritten as

$$P(Y_{ik}|\theta_i,a_k,b_k)=\frac{e^{d(a_k\theta_i-b_k)}}{1+e^{d(a_k\theta_i-b_k)}}=\left(1+e^{d(b_k-a_k\theta_i)}\right)^{-1} \quad .$$

2.1.4.2 Three-parameter model

17 θ. 16 When latent variable refers to a person characteristic, as an

The normal ogive model can be adapted through the aggregation of other parameter, which will describe the lower asymptote of the ICC:

$$\begin{aligned} P(Y_{ik}=1|\theta_i,a_k,b_k,c_k) &= c_k+(1-c_k)\Phi(a_k\theta_i-b_k) \\ &= \Phi(a_k\theta_i-b_k)+c_k(1-\Phi(a_k\theta_i-b_k)) \quad , \end{aligned} \tag{2.5}$$

where c_k is the "guessing" parameter of the $k-$th item. Thus, the correct response probability will be given by the guessing parameter c_k plus a success probability, scaled by the c_k complement, which is the second term. The expression at the bottom of equation (2.5) is the expansion and c_k factorization. Then, the logistic probability is

$$\begin{aligned} P(Y_{ik} = 1|\theta_i, a_k, b_k, c_k) &= c_k + \frac{1 - c_k}{1 + e^{b_k - a_k\theta_i}} \\ &= \frac{1}{1 + e^{b_k - a_k\theta_i}} + \frac{c_k}{1 + e^{a_k\theta_i - b_k}} \quad . \end{aligned} \tag{2.6}$$

Note that when $c_k = 0$, the three parameter model is equal to the two parameter Rasch model, since two parameter model and three parameter model differ in a constant which existence in the model depends only of the value of c_k. Then this changes the interpretation for models with $c_k > 0$, because in the three parameter model, the proportion of a correct response $\frac{b_k}{a_k} = 0.5 + c_k$, while in the two parameter model, the same probability is equal to 0.5.

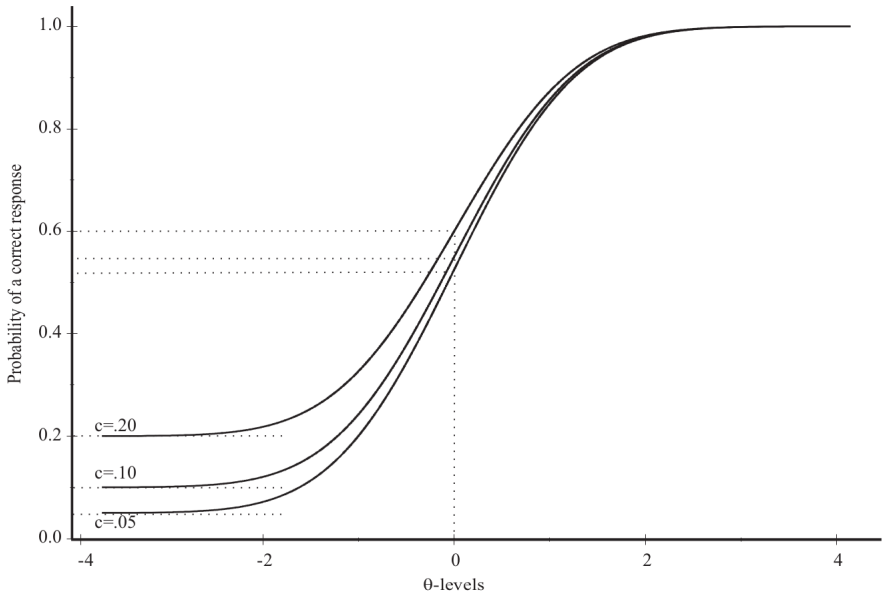

Figure 2.3: ICC for three curves related to three different guessing levels, from (Fox, 2010).

This model can be plotted with different c_k levels, where three ICCs have equal discrimina-

tion and difficulty level, but different guessing levels. This situation is represented graphically in Figure 2.3, where guessing levels are low (0.05), medium (0.1) and high (0.2). Then, the probability of a correct response when guessing, is the height of the lower asymptote. Since in the Figure 2.3, can be easily seen that for high ability respondents, the guessing effect tends to zero, that is, in bottom-right side of equation 2.6, when θ_i is big enough, $\frac{c_k}{1+e^{a_k\theta_i - b_k}} \downarrow 0$.

2.1.5 Multi-faceted Rasch models

Commonly, is not only important to know if an answer is good or bad, instead, is interesting to know "how good" or "how bad" it is. Also, Likert scale items are commonly used in surveys, where the problem is not analyze a binary variable. When the response item can be ordered with respect to the ability parameter, these are known as graded responses. This kind of model contains more parameters, but is possible to capture more precise information about the ability parameter θ_i, when the extension to more categories is used.

Cohen (1983) found that statistical information is loss through dichotomization, thus it increases when comparing with polytomous data. Polytomous response data, for better understanding purposes, is also called ordinal response data.

In order to develop the intuition, consider as example an english test, where relevant facets are already identified, such as tasks and raters, the multi-faceted Rasch model (MFRM) can be expressed as (Eckes, 2015):

$$\log\left(\frac{p_{nljk}}{p_{nljk-1}}\right) = \theta_n - \delta_l - \alpha_j - \tau_k \quad , \tag{2.7}$$

where

$$
\begin{aligned}
p_{nljk} &= \text{probability of } n-\text{th examinee to recieve a rating } k \text{ from rater } j \text{ on task } l,\\
p_{nljk-1} &= \text{probability of } n-\text{th examinee to recieve a rating } k-1 \text{ from } j-\text{th rater at } l-\text{th task},\\
\theta_n &= \text{ ability of } n-\text{th examinee}\\
\delta_l &= \text{difficulty of } l-\text{th task},\\
\alpha_j &= \text{severity of } j-\text{th rater},\\
\tau_k &= \text{difficulty of receiving a rating } k, \text{ relative to } k-1.
\end{aligned}
$$

Masters (1982) developed an unidimensional latent trait model for responses scores for more of one category, called partial credit model (PCM). The model requires many steps, where each step is completed through a partial credit. The probability of a response in a category c, with $c \in \{1, \ldots, C_K\}$ of the $k-$th item is given by:

$$
P(Y_{ik} = c|\theta_i, \boldsymbol{\kappa}_k) = \frac{e^{\sum_{l=1}^{c}(\theta_i - \kappa_{k,l}}}{\sum_{r=1}^{K}\left(e^{\sum_{l=1}^{r}\left(\theta_i - \kappa_{k,l}\right)}\right)} \quad ,
$$

where $\kappa_{k,l}$ is the item step difficulty parameter, and $\sum_{l=1}^{1}(\theta_i - \kappa_{k,l}) \equiv 0$. Note that items can have different number of categories between each other. Given that each item parameter is locally defined the set of items are not constraint to a specific order, because as the local item parameter is defined respect two adjacent categories (the previous, and the next one). Instead of taking all categories simultaneously, since this will be detrimental for the model because it would need all categories order with the same length. The PCM was developed by Muraki (1993, 1992); this model allows items to have different slopes parameters for different items.

It is not possible to directly model through cumulative probability in the PCM, but there is a result of summing the categorical response functions. Samejima (1997) introduced the graded re-

sponse model, where the cumulative probabilities are directly modeled. The probability of response in a category $c-1$ (or above) minus the probability of responding in category c (or above). Denoting K as the number of response categories of the current $i-$th item, there are $K-1$ thresholds between the response option set. This can be mathematically expressed as

$$\begin{aligned} P(Y_{ik}=c|\theta_i,\boldsymbol{\kappa}_k) &= P(Y_{ik}\geq c-1|\theta_i,\boldsymbol{\kappa}_k)-P(Y_{ik}\geq c|\theta_i,\boldsymbol{\kappa}_k) \\ &= \int_{\kappa_{k,c-1}}^{\infty}\psi(z;a_k\theta_i)dz-\int_{\kappa_{k,c}}\psi(z;a_k\theta_i)dz \\ &= \Psi(a_k\theta_i-\kappa_{k,c-1})-\Psi(a_k\theta_i-\kappa_{k,c}) \\ &= \frac{e^{a_k\theta_i-\kappa_{k,c-1}}}{1+e^{a_k\theta_i-\kappa_{k,c-1}}}-\frac{e^{a_k\theta_i-\kappa_{k,c}}}{1+e^{a_k\theta_i-\kappa_{k,c}}} \quad . \end{aligned} \tag{2.8}$$

This is known as the logistic distribution. The inverse logit function is defined as

$$\text{logit}^{-1}(x)=(1+e^{-x})^{-1} \quad .$$

This is equal to the cumulative distribution function of the logistic distribution. Threshold can be defined as $\boldsymbol{\beta'}=(\beta_1,\ldots,\beta_{K-1})\in\mathbb{R}^{K-1}$, such that $\beta_k<\beta_{k+1}$, and $\eta\in\mathbb{R}$. Also, $K\in\mathbb{N}$ with $K>2$. Thus, for $k\in\{1,\ldots,K\}$ the probabilistic mass function (pmf) of the ordered logistic distribution can be defined as

$$g(k|\eta,\boldsymbol{\beta})=\begin{cases} 1-\text{logit}^{-1}(\eta-\beta_1) & \text{if } k=1 \quad , \\ \text{logit}^{-1}(\eta-\beta_{k-1})-\text{logit}^{-1}(\eta-\beta_k) & \text{if } 1<k<K \quad , \\ \text{logit}^{-1}(\eta-\beta_{K-1}) & \text{if } k=K \quad . \end{cases} \tag{2.9}$$

When setting $\beta_0=-\infty$ and $\beta_K=\infty$, $\text{logit}^{-1}(-\infty)=0$ and $\text{logit}^{-1}(\infty)=1$. This work only for edge cases ($k=\{1,K\}$).

For classical logistic regression, the predictors are the values taken by η, and β_k-values are regression parameters, which are estimated. The multi-faceted models are item response models. Thus, the predictor values have to be estimated, since they are latent variables; β_k values are item parameters. Already knowing that the η's are predictors, note that if $\beta_k<\eta\leq\beta_k$, then the $k-$th

category has the higher probability, this works since η and β_k are in the same parametric space. This can be visualized in Figure 2.4, where $\beta-$parameters are cutting points, which determine the upper and lower bounds for the $k-$th category, depending of the value of η.

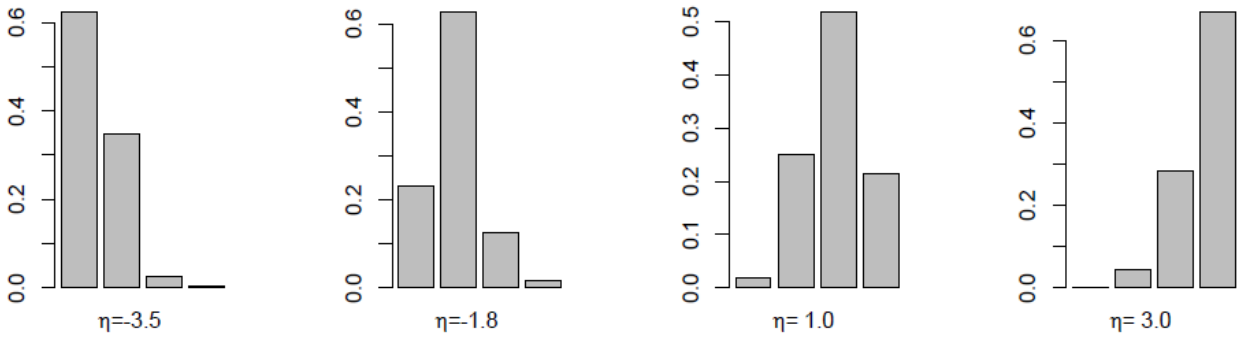

Figure 2.4: Probability mass function of the ordered logistic distribution for different values of η, with $\boldsymbol{\beta} = (-3.0, 0.0, 2.3)^t, K = 4$.

2.2 Bayesian multi-faceted model for measuring professor performance

In order to analyze SET data, the bayesian multi-faceted (BMF) model is proposed to be applied. It is also possible to apply this kind of model for other areas, because BMF model could be used in any application of the MFRM.

As was described before, latent variables play an important role to measure the performance of examinee, in this case, the latent variable is the professor performance; that is, a variable which never can be measured directly. In statistical terms, latent variables are considered as random effects (Bartholomew et al., 2011). Also, the severity or leniency of the student is treated as latent variable, which is not either directly measured from questionnaires. For every course, each professor is qualified for all the items by each student. An important assumption is that each item is based on a Likert scale with K_j categories. The questionnaires has p items, and N evaluated professors.

To grade the professor, the student selects a category of the item. Then, the BMF model is defined as follows: Let θ_i be the latent variable which measures the performance of the $i-$th professor. Let $\boldsymbol{\gamma}_i' = (\gamma_{i1}, \ldots, \gamma_{i,n_i})$ be the vector of latent variables which measures the severity of the students evaluating professor i. The value n_i is the total number of students evaluating professor i. Let $\boldsymbol{\beta}_j' = (\beta_{j1}, \ldots, \beta_{jK_j-1})$ be the cut points of item j associated to each response category. Let Y_{ijs} be the rating that the $is-$th student assigns to professor i for item j. Thus, the conditional probability that $[Y_{ijs} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j]$ is given by

$$Prob[Y_{ijs} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j] = logit^{-1}(\eta_{is} - \beta_{j(k-1)}) - logit^{-1}(\eta_{is} - \beta_{jk}) \quad , \tag{2.10}$$

where $k = 1, \ldots, K$, $\eta_{is} = \theta_i - \gamma_{is}$. This model is a direct implication from the logistic distribution. A difference in this model, is the fact that it is not identifiable, as it is common in response models. As was mentioned before, the performance is assumed as a latent variable with normal density.

For this case, the professor performance will be assumed as a standard normal distribution, that is, $\theta_i \sim N(0, 1)$. The following prior distributions are assigned:

$$\begin{aligned}
\gamma_{is} &\sim N(0, \sigma_\gamma^2) \quad , \\
\beta_{jk} &\sim N(\mu_\beta, \sigma_\beta^2) \quad , \text{ cut points} \\
\mu_\beta &\sim N(0, 2) \quad , \\
\sigma_\gamma &\sim Cauchy(0, 2)I_{(0,\infty)} \quad , \\
\sigma_\beta &\sim Cauchy(0, 2)I_{(0,\infty)} \quad .
\end{aligned} \tag{2.11}$$

There are some controversial assumptions, which allow to have the posterior distribution. First one is that responses of the students are independent between each other. The Second, is that the response of a student to a professor are independent for each other. The first assumption may be violated if a student has more than one class with a professor, for example if the student take

two courses with the same professor. However, since the subjects are different for this assumption, to violate the first assumption is not detrimental for posterior distribution development. Moreover, there is not way to know when this happens because the student responses are anonymous. On the other hand, every item in the questionnaire is designed to measure a different aspect of the professor in the classroom. Therefore, independent answers are expected by students for each question.

Let $p_{ijsk} = Prob[Y_{ijs} = k|\theta_i, \gamma_{is}, \boldsymbol{\beta}_j]$. Therefore, the posterior distribution for the Bayesian multi-faceted model will be given by

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y}) \propto \prod_{i=1}^{N}\prod_{j=1}^{p}\prod_{s=1}^{n_i}\prod_{k=1}^{K_j}[p_{ijsk}]^{\chi_k(y_{ijs})} \times p(\theta_i)p(\beta_{jk}|\mu_\beta, \sigma_\beta)p(\gamma_{is}|\sigma_\gamma)p(\mu_\beta)p(\sigma_\beta)p(\sigma_\gamma) \quad ,$$

Bold greek letters are complete vector parameters, the $\boldsymbol{y}-$vector represents the complete vector of observed responses, and $\chi_k(y_{ijs})$ is given by

$$\chi_k(y_{ijs}) = \begin{cases} 1, \text{ if } y_{ijs} = k, \\ 0, \text{ otherwise.} \end{cases} \tag{2.12}$$

The subject relevance can be modeled similarly as professor performance, but with different latent variables. For this case, $\boldsymbol{\theta}$ is the vector of subjects' relevance measures, and they are measured by the items. Though the students are the same, the student severity measure change in every case.

2.3 Estimation of BMF models

Carpenter et al. (2017) developed a probabilistic programming language called STAN, which is useful to simulate very complex bayesian models of high performance. It has an easy programming syntax, as can be seen in Appendix A. There are interfaces which allow to compile Stan code, such as Python, R, Julia, etc (RStan, 2017). RStan codes are used to the IRT analysis. Stan

uses a just in time compiler; this implies that at the first time the program is runned, the code is compiled. A natural question is, how does the Stan compiler work? It translates the Stan code to C++, and compiles the C++ code. Stan uses automatic differentiation procedure to implement the required lagrangian operations (Bucker et al., 2006). Stan codes implement Hamiltonian Monte Carlo samplers (Neal, 2011), and the No-U-turn sampler (Hoffman and Gelman, 2014). For more information about the Stan implementation, see Luo and Jiao (2017), who explain how to use Stan codes for IRT cases.

The original data set is structured as an $M \times (p + 2)$ matrix, where M is the number of surveys and p is the number of elements in the survey. 1 to p columns contain students responses assigned to professor to each item. These scores are between 1 and K_j, for each column. Professor *id* column corresponds to the $p+1$ column, and the student *id* is the $p+2$ column. Since students are anonymous, students *id* are generated in a previous step. The student *id* is useful to work with for complementary analysis, when estimating student severity.

In order to avoid missing values, the re-structured data is a $N \times 4$ matrix, where N is the number of total answers. For professor related questionnaires, the first column is the answer value, second column is the professor *id*, third column is the item *id* of the current response, and the fourth column is the student *id*.

Note that missing values are not a problem for this data structure, because they are omitted. In Appendix A, the Stan code for measuring professor performance is available. For measuring subject relevance, the only difference is in 15^{th} line, where the latent trait is the subject relevance. The same code is available as supplementary material.

Main elements in a Stan program are shown in Appendix A. Stan programs are splitted into blocks. These blocks are *data, transformed data, parameters, transformed parameters, model*, and

generate quantities. For the appendix code, it is not a requirement to have *transformed data* and *transformed parameters*.

The block *data* in Appendix A defines the input data. The key words *lower* and *upper* are used to define constraints in the data, that are verified at reading time. For example, in line 4, the minimum value is 10. In section *parameters*, estimated parameters are defined. It is also possible to make restrictions. Observe in line 20, as σ_γ is defined as positive. In this case, restrictions are considered at sampling time. Stan transforms automatically the constrained variables, in such a way that the internal variables has unconstrained domain. An important step is to introduce and add the lagragiang factors to the log-posterior distribution, which is required for transformations. This is automatically done. The procedure is equivalent for *transformed parameters*. The bayesian model is explicit in a declarative way in the *model* block. This block is similar in BUGS Spiegelhalter et al. (1997). Moreover, at *generate quantities* block, some derivate quantities can be computed. This is a useful block, because any functional can be calculated through it; such as goodness of fit statistic. In this block was calculated the mean of the $\beta-$parameters for each item.

The *rstan* library allows to Appendix D illustrates how to run a stan code from R language (RStan, 2017). Line 4 is required to detect cores of inside the computer. In line 22 the program is called to run only 4 iterations, this is to compile the code. After that, in line 24, the compiled code is used by Stan to run the procedure.

2.4 Evaluating model adequacy tools for BMF models

There are some necessary assumptions to ensure that BMF models are good to measure professors' performance, such as:

(1) **Unidimensionality**. The items in the questionnary are designed to measure a unique

construct.

(2) **Independence**. The latent traits θ_i must be independent from the items and from the socio-demographic characteristics of the students, their severity and their score in the course.

(3) **Goodness of fit**. It is necessary to assess the quality of the model to fit the data.

For this thesis, unidimensionality is assumed for BMF model. This means that the items from the questionnaire are designed to measure a unique construct. Thus, the first thing to do is to verify the data set from the survey is really unidimensional. In the next subsection, some tools are introduced to verify the unidimensionality of the data.

Second important assumption to verify is that latent traits θ_i are independent from the questionnaire, the students' socio-demographic characteristics, and severity, as well as their scores in the course. An important assumption is that experts, who design the questionnaires, guarantee the independence between questionnaires and professors. This concern has theoretical aspects involved, which are based on IRT (Bock, 1997; Baker and Seok-Ho, 2004; Birnbaum, 1968). Nevertheless, the independence must be assessed between latent traits and socio-demographic characteristics of the students, their severity and their grade in the course. Tools to evaluate independence in BFM models are introduced in 2.4.2 and 2.4.3. On the other hand, in item response models, is impractical to assess the goodness of fit of the model to the complete data. Instead, it is assessed the adequacy of the model for fitting the data of each one of the examinees, and the data of each one of the items. In section 2.4.5 we propose original GoF statistics to do that.

2.4.1 Analysis of unidimensionality

The main principle to define a scale to measure as an unique construct is that the resulting data be unidimensional. To perform the data analysis, Apodaca and Grad (2005) used factorial techniques to do it, and to review data dimensions. Unidimensionality must be the first assumption

to be verified.

There are several tools to evaluate dimensionality of data. Three of them are used for this study: principal component analysis (PCA) (Jollife, 2002), Cronbach's alpha coefficient (Cronbach, 1951; Lord and Novick, 2013), and Cronbach Mesbah curve (CMC) (Cameletti and Caviezel, 2015). Cronbach's alpha reliability is an index between 0 and 1, which is based on the variance of every score item. When this value is 0, it means the items are not correlated between each other; and when the index is equal to 1, it means all the items are the same. According to experts, values greater than 0.75 or 0.80 means the data is measuring a unique construct. For a more detailed explanation of this, see (Lord and Novick, 2013).

There are different interpretations about data dimension based on the plot of the eigenvalues. In general, a unique large eigenvalue compare to the other eigenvalues suggests unidimensionality. From this perspective, the left panel of Figure 5.1 suggests that the data seems unidimensional. On the other hand, the right panel of the figure shows the CMC. This curve shows the Cronbach's alpha coefficient after taken out the item for which the preserved data has maximal Cronbach's alpha coefficient. If the data is unidimensional, the curve is monotonically increasing. For details see Cameletti and Caviezel (2015).

2.4.1.1 χ^2-distance

To make a comparison between two distributions through its data, the Chi-square (χ^2) intuition is used to develop a distance measure between different random vectors. To develop the main idea of this distance, consider χ^2-statistic:

$$\chi^2 = \sum_{i \in \text{cells}} \frac{(O_i - E_i)^2}{E_i} \quad .$$

For contingency tables, the χ^2-statistics is the difference of the observed data and its expectation, over its expectation. Actually, the χ^2-distance conserve this intuition, and it is highly related to

the weighted euclidean distance (Greenacre, 2017). This distance is commonly used to compare a big number of multivariate data distributions (Daliri, 2013).

The formula of the distance is described as follows:

$$\chi^2(l,k) = \sqrt{\sum_j \frac{1}{x_{+j}} \left(\frac{x_{lj}}{x_{l+}} - \frac{x_{kj}}{x_{k+}} \right)^2} \quad , \tag{2.13}$$

where $x_{+j} = \sum_i x_{ij}$ y $x_{i+} = \sum_j x_{ij}$. For more information about this, Greenacre (2017) has an entire chapter related to this measure.

2.4.2 Kendall correlation

Consider the fact that a SET data set is ordinal, not continuous; also, the associations between these ordinal variables in the model are not necessarily linear. Thus, the data analysis needs to involve a more realistic association measure than Pearson's correlation. Spearman's correlation (Spearman, 1904) is a often used measure to find associations between ordinal variables, but is a controversial measure. Spearman's correlation is based on the assumption that ranks are continuous variables and that Pearson's correlation based on ranks is also valid. Kendall's rank correlation coefficient 5.1, which is commonly referred to as Kendall's tau (τ) coefficient, is highly recommended for these cases. From a bayesian point of view, Kendall's tau is computed for each sample for the estimation process.

2.4.3 Mutual information

There are many useful coefficients to measure independence. Mutual information (MI) is singled out by its theoretical information background (Cover and Thomas, 1991; Ash, 1990). MI is zero if and only if both compared random variables are strictly independent. A characteristic of MI, is that the calculus of it may be complicated. The package *infotheo* of R Meyer (2015) is used for this thesis to generate MI for every bayesian iteration. As usually happens when having a set of N bivariate measurements, $z_i = (x_i, y_i)$, for $i = 1, \ldots, N$, which commonly are assumed as

independent identically distributes (iid) realizations of a random variable $Z = (X, Y)$ with density $f(x, y)$, where x and y can be scalars or higher dimensional elements. Although densities with singularities can be allowed, it will be assumed that the density is a proper smooth function. Thus, the integrals bellow must exist somehow. In particular, it will be always assumed that $0 \log(0) = 0$, i.e. it is not necessary to assume that densities are strictly positive. The marginal densities of X and Y are $f_x(x) = \int f(x, y)dy$ and $f_y(y) = \int f(x, y)dx$. The MI is defined as

$$I(X, Y) = \int \int f(x, y) \log \frac{f(x, y)}{f_x(x) f_y(y)} dxdy$$

The units in which information is based determines the logarithm base. For example, when the base of the logarithm is 2, it is equal to the information measured in bits. But following with the current SET problem, natural logarithm will be used, since the information is measured in the natural set.

2.4.4 Normalized Mutual Information

In order to have a measure between 0 and 1, a variation of version of normalized mutual information have been introduced (Kavalseth, 2017). The entropy or uncertainty of a random variable X is defined as

$$H(X) = -\int f_x(x) \log f_x(x) dx.$$

Then, the normalized mutual information (NMI) which will be used, is be given by

$$\kappa(X, Y) = \frac{I(X, Y)}{\sqrt{H(X)H(Y)}} \quad . \tag{2.14}$$

2.4.5 Goodness of fit statistics

In this subsection, to simplify the notation easier, it is assumed $\boldsymbol{\theta}$ as the complete vector parameter, and the observed data is $\boldsymbol{y}$. The likelihood of the observations is $f(\boldsymbol{y}|\boldsymbol{\theta})$ and the prior

density of $\boldsymbol{\theta}$ is $\pi(\boldsymbol{\theta})$, from where the model specification is $f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$. The predictive distribution of unobserved values $\boldsymbol{\omega}$ is denoted by $f(\boldsymbol{\omega}|\boldsymbol{z})$, and defined as

$$f(\boldsymbol{\omega}|\boldsymbol{y}) = \int f(\boldsymbol{\omega}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}.$$

For the goodness of fit (GoF) procedure, the data $\boldsymbol{\omega}$ is called as replicated data, and the associated random variable will be defined as $\boldsymbol{W}$. That is, $\boldsymbol{W}$ has a density of the form $f(\boldsymbol{\omega}|\boldsymbol{y})$. A highlighted proposal is which Box did, that consists on computing the expectation $E[g(\boldsymbol{W},\boldsymbol{\theta})|\boldsymbol{y}]$ of the set of relevant statistic $g(\boldsymbol{W},\boldsymbol{\theta})$. Box (1980), Gelfand et al. (1992) based in discrepancy measures a variety of checking functions. This work uses a checking function based on discrepancy statistic $D(\boldsymbol{W},\boldsymbol{\theta}) = -2\log f(\boldsymbol{W}|\boldsymbol{\theta})$. Gelman et al. (1996) develops the relevance and importance of using discrepancy measures. The constant 2 in the statistic is obviously redundant. The formula keeps it to have a counterpart to the log likelihood statistics commonly used in frequentist statistics.

Using Box (1980) and Gelfand et al. (1992) notation for discrepancy statistic D as Gelman et al. (1996) proposes, the goodness of fit statistics is written as follows. Define as $B_{\omega,\theta} = \{(\boldsymbol{\omega},\boldsymbol{\theta}) : D(\boldsymbol{\omega},\boldsymbol{\theta}) \leq D(\boldsymbol{y},\boldsymbol{\theta})\}$. The test statistic is defined as $g(\boldsymbol{W},\boldsymbol{\theta}) = I_{B_{\omega,\theta}}(\boldsymbol{W},\boldsymbol{\theta})$. The statistical decision is based on the value $d_{\omega,\theta}$, which is

$$\begin{aligned} d_{\omega,\theta} &= P(B_{\omega,\theta}) \\ &= E[g(\boldsymbol{W},\boldsymbol{\theta})|\boldsymbol{y}] \\ &= \int\int g(\boldsymbol{\omega},\boldsymbol{\theta})f(\boldsymbol{\omega}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta}d\boldsymbol{\omega} \quad . \end{aligned}$$

The interpretation of $d_{\omega,\theta}$ is relatively easy. For values around 0.5, it means the model fits the data properly. A reasonable range can be between 0.05 and 0.95, or being more strict, between 0.1 and 0.9.

2.4.6 Goodness of fit for bayesian multi-faceted models

I item response models, it is important to fit separately the model to items and people. Consequently, it is used the complete data to the adequacy of the model; that is, the data of each evaluated professor and each item. Let $\boldsymbol{y}_{i\cdot\cdot}$ be the complete vector of responses to professor i, $\boldsymbol{y}_{\cdot j\cdot}$ be the complete vector of responses to the $j-$th item. The corresponding replicated data will be denoted as $\boldsymbol{\omega}_{i\cdot\cdot}$, and $\boldsymbol{\omega}_{\cdot j\cdot}$ respectively. To be sure the GoF of the model to the $i-$th professor's data, it will be used the predictive density, which is given by

$$f(\boldsymbol{\omega}_{i\cdot\cdot}|\boldsymbol{y}_{i\cdot\cdot}) = \int f(\boldsymbol{\omega}_{i\cdot\cdot}|\theta_i, \boldsymbol{\gamma}_i, \boldsymbol{\beta})\pi(\theta_i, \boldsymbol{\gamma}_i, \boldsymbol{\beta}|\boldsymbol{y}_{i\cdot\cdot})d\theta_i d\boldsymbol{\gamma}_i d\boldsymbol{\beta} \quad . \tag{2.15}$$

Similarly, to asses the fit of the model to the $i-$th item is

$$f(\boldsymbol{\omega}_{\cdot j\cdot}|\boldsymbol{y}_{\cdot j\cdot}) = \int f(\boldsymbol{\omega}_{\cdot j\cdot}|\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}_j)\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}_j|\boldsymbol{y}_{\cdot j\cdot})d\boldsymbol{\theta} d\boldsymbol{\gamma} d\boldsymbol{\beta}_j \quad .$$

2.4.6.1 Computational approach to compute the GoF statistic

A complete $t-$th sample from the posterior distribution is defined as $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \sim \pi(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{y})$, for $t = 1, \ldots, T$. The Stan code generates these sample data from the estimation output. From the $i-$th professor, the GoF statistic can be calculated following then ext procedure:

(1) Obtain a replicate data vector $\{\boldsymbol{\omega}_{ijs}; j = 1, \cdots, p; s = 1, \cdots, n_i\}$ from the ordered logistic distributions given by

$$g_{ijs}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \boldsymbol{\beta}_j^{(t)}) = \text{logit}^{-1}(\theta_i^{(t)} - \gamma_{is}^{(t)} - \beta_{j,k-1}^{(t)}) - \text{logit}^{-1}(\theta_i^{(t)} - \gamma_{is}^{(t)} - \beta_{j,k}^{(t)}),$$

$k = 1, \cdots, K_j$. Remember that $\beta_{j0} = -\infty$, and $\beta_{jK_j} = \infty$.

(2) The discrepancy measure for $\boldsymbol{\omega}_{i\cdot\cdot}$ is given by

$$D_i(\boldsymbol{\omega}_{i\cdot\cdot}|\theta_i^{(t)}, \boldsymbol{\gamma}_i^{(t)}, \boldsymbol{\beta}^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(\omega_{ijs})\log g_{ijs}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \boldsymbol{\beta}_j^{(t)}).$$

(3) The discrepancy measure for $\boldsymbol{y}_{i\cdot\cdot}$ is given by

$$D_i(\boldsymbol{y}_{i\cdot\cdot}|\theta_i^{(t)}, \boldsymbol{\gamma}_i^{(t)}, \boldsymbol{\beta}^{(t)}) = -2\sum_{j=1}^{p}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(y_{ijs})\log g_{ijs}(k|\theta_i^{(t)}, \gamma_{is}^{(t)}, \boldsymbol{\beta}_j^{(t)}).$$

(4) The GoF statistic is given by

$$p_i = \frac{1}{T}\sum_{t=1}^{T} 1_{(D_i(\boldsymbol{\omega}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\beta}^{(t)})-D_i(\boldsymbol{y}_{i..}|\theta_i^{(t)},\boldsymbol{\gamma}_i^{(t)},\boldsymbol{\beta}^{(t)})>0)} \quad .$$

When p_i values are close to 0.5, it imply a good fit. The $j-$th item of the data, the GoF statistic can be computed similarly. For this, the discrepancy measure for $\boldsymbol{\omega}_{.j.}$ will be

$$D_j(\boldsymbol{\omega}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)}) = -2\sum_{i=1}^{N}\sum_{s=1}^{n_i}\sum_{k=1}^{K_j}\chi_k(y_{ijs})\log g_{ijs}(k|\theta_i^{(t)},\gamma_{is}^{(t)},\boldsymbol{\beta}_j^{(t)}) \quad ,$$

then, the GoF statistic has the next expression:

$$p_j = \frac{1}{T}\sum_{t=1}^{T} 1_{(D_j(\boldsymbol{\omega}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)})-D_j(\boldsymbol{y}_{.j.}|\boldsymbol{\theta}^{(t)},\boldsymbol{\gamma}^{(t)},\boldsymbol{\beta}_j^{(t)}))}.$$

Chapter 3

Latent Dirichlet Allocation

Latent Dirichlet allocation is a generative process from a set of documents. This assumes that in a given corpus, each document has a topic distribution, such that each topic has a distribution over terms whose belong to a corpus. The generative model simulates how documents are produced; the main idea behind this, is that documents can be represented as random mixtures of hidden latent variables.

LDA can be explained in the following way: suppose that the objective is to study if a word appears or not inside a corpus, thus this would be a Bernoulli trial; then, it could be of interest the counting of a particular word into a corpus, that would be a binomial process. But when considering the counting of many words, such as is the case of topic modeling, the whole set filled by terms will have a multinomial distribution. Note that a beta distribution is the conjugate prior of the binomial distribution; then, the vectorial extension of this conjugate distribution is the Dirichlet distribution.

More formally, in a document $\mathbf{w} = (w_1 \ldots, w_N)$, which contains N words out of V different terms, $w_i \in \{1, \ldots, V\}$ for $i = 1, \ldots, N$ inside a corpus $\mathbf{D} = \{\mathbf{w}_1, \ldots, \mathbf{w}_M\}$, the LDA model can be expressed through the next procedure:

(1) The term distribution $\boldsymbol{\beta}$ for each topic, $p(\boldsymbol{w}_n|z_n, \boldsymbol{\beta})$ is determined by:

$$\boldsymbol{\beta} \sim Dirichlet(\boldsymbol{\delta}) \quad . \tag{3.1}$$

(2) The topic distribution proportions $\boldsymbol{\theta}$ for each $\mathbf{w}-$document, will by determined by:

$$\boldsymbol{\theta} \sim Dirichlet(\boldsymbol{\alpha}) \quad . \tag{3.2}$$

(3) Then, for each of the N words w_i:

(a) Choose a topic $z_i \sim Multinomial(\boldsymbol{\theta})$.

(b) Choose a term w_i from $p(w_i|z_i, \boldsymbol{\beta})$.

For given α and β, the joint distribution of the topic mixture θ, a set of topics $\mathbf{z}$, and a set of N words $\mathbf{w}$, will be given by:

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \boldsymbol{\beta}) \quad , \tag{3.3}$$

3.1 Parameter estimation

LDA uses maximum likelihood estimation of the parameters for the model; the estimation is made thorugh the maximization of α and β parameters. For each document $\mathbf{w} \in \mathbf{D}$, the log-likelihood will be given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \ln p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}) \quad , \\ &= \ln \int \left\{ \sum_{\mathbf{z}} \left[\prod_{i=1}^{N} p(w_i|z_i, \boldsymbol{\beta}) p(z_i|\boldsymbol{\theta}) \right] \right\} p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} \quad . \end{aligned} \tag{3.4}$$

Note that the function $\ln p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ can not be expressed in a closed form in order to be computed. Thus, it is necessary a variation of the expectation maximization (EM) algorithm (VEM). Note that in this case it is in the presence of unknown latent variables $\boldsymbol{\theta}$ and $\mathbf{z}$; thus it makes sence the idea of applying the algorithm for this case. The use of the VEM allows to use another density instead of $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$, which is replaced by the variational density function $q(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ by $q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$. This means that $\mathrm{E}_p\{\ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\}$ is replaced by $\mathrm{E}_q\{\ln p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\}$. The

parameter variational distribution can variate through the document, which is not the case for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For a document $\mathbf{w}$, the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\phi}$ are determined by:

$$(\boldsymbol{\gamma}^*, \boldsymbol{\phi}^*) = \arg\min_{(\boldsymbol{\gamma},\boldsymbol{\phi})} D_{KL}\{q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})\} \quad , \tag{3.5}$$

where D_{KL} is the Kullback-Leiber (KL) divergence between variational distribution and the truth posterior $p(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})$. Thus, the variational distribution will be given by:

$$q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi}) = q_1(\boldsymbol{\theta}|\boldsymbol{\gamma}) \prod_{i=1}^{N} q_2(z_i|\phi_i) \quad , \tag{3.6}$$

where q_1 is a Dirichlet distribution with parameters $\boldsymbol{\gamma}$, and q_2 is the multinomial distribution with ϕ_i parameters. The work of Blei et al. (2003) shows that variational parameters $\ln p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ are equivalent to

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) + \arg\min_{(\boldsymbol{\gamma},\boldsymbol{\phi})} D_{KL}\{q(\boldsymbol{\theta}, \mathbf{z}|\boldsymbol{\gamma}, \boldsymbol{\phi})||p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \boldsymbol{\alpha}, \ \boldsymbol{beta})\} \quad , \tag{3.7}$$

where

$$\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = \mathrm{E}_q\{\ln p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})\} - \mathrm{E}_q\{\ln q(\boldsymbol{\theta}, \mathbf{z})\} \quad . \tag{3.8}$$

This works because maximizing the lower bound of $\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \boldsymbol{\beta})$ is equivalent to minimize the KL divergence between the variational posterior distribution and the real one. For estimation purposes, the EM algorithm is applied until the convergence of log-likelihood lower bound. This can be made through the following procedure:

E-step : For each document, find the optimizing value of variational parameters $\boldsymbol{\gamma}$, and $\boldsymbol{\phi}$.

M-step : Maximize the log-likelihood lower bound respect the model parameters $\boldsymbol{\alpha}$ $\boldsymbol{\beta}$.

The EM algorithm is used to study latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. That is, the variational posterior distribution is a good approximation to the real posterior distribution. Gibbs sampling is used to generate samples from this posterior.

3.2 Gibbs sampling

To work directly with $\ln p(\mathbf{w}|\boldsymbol{\alpha}, \boldsymbol{\beta})$ may be very complex; for this case, a Monte Carlo method was used in its place. Gibbs sampling is a particular case of Markov Chain Monte Carlo (MCMC). Thus, considering the Metropolis-Hastings algorithm (Robert and Casella, 2010), which basically works with every iteration for a univariate distribution with a posterior, as follows:

(1) Compute

$$r(\theta_{\text{new}}, \theta_{t-1}) = \frac{\text{Posterior}(\theta_{\text{new}})}{\text{Posterior}(\theta_{t-1})} \quad . \tag{3.9}$$

(2) Calculate

$$\zeta(\theta_{\text{new}}, \theta_{t-1}) = \min\{r(\theta_{\text{new}}, \theta_{t-1}), 1\} \quad . \tag{3.10}$$

(3) Generate $u \sim Uniform(0, 1)$.

(4) If $u < \zeta(\theta_{\text{new}}, \theta_{t-1})$, then $\theta_t = \theta_{\text{new}}$; in the contrary case $\theta_t = \theta_{t-1}$.

For a multivariate distribution, such as in the current case, the procedure is analogue; to obtain the posterior distribution of the mixed model, $\mathbf{w}$ and $\mathbf{z}$ are used as all the terms and assignment vectors from the entire corpus. The topic selection of a particular term depends of the current topic assignment of all positions.

Having k topics, the $i-$th term probability in a given document, can be written as:

$$p(w_i) = \sum_{K=1}^{k} p(w_i|z_i = K)p(z_i = K) \quad , \tag{3.11}$$

where z_i is a latent variable which indicates the $i-$th topic assignment. Intuitively $p(\mathbf{w}|\mathbf{z})$ will contain more information about the words occurrences than $p(\mathbf{z})$.

The conditional posterior distribution will be given by:

$$\begin{aligned} p(z_i = K|z_{-i}, \mathrm{w}) \propto \quad & p(w_i|z_i = K, z_{-i}, w_{-i}) \\ \times \quad & p(z_i = K|z_{-i}) \quad , \end{aligned} \tag{3.12}$$

where z_i is assigned to all the topics except the $i-$th one. This is done through the direct application of the Bayes theorem property, because $p(w_i|z_i = K, z_{-i}, w_{-i})$ is a likelihood, and $p(z_i = K|z_{-i})$ is a probability. The first term of the proportion $p(w_i|z_i = K)$ can be obtained as the integral:

$$\int p(w_i|z_i = K, \phi^{(K)})p(\phi^{(K)}|\mathbf{z}_{-i}, \mathbf{w}_{-i})d\phi^{(K)} \quad . \tag{3.13}$$

Here, $\phi^{(K)}$ is the multinomial distribution associated to the $K-$th topic, and the integral is a functional over all the distributions. The second term of the integral can be found using one of the properties of the Bayes theorem, again:

$$p(\phi^{(K)}|\mathbf{z}_{-i}, \mathbf{w}_{-i}) \propto p(\mathbf{w}_{-i}|\phi^{(K)}, \mathbf{z}_{-i})p(\phi^{(K)}) \quad . \tag{3.14}$$

Since it is known that $p(\phi^{(K)}) \sim \text{Dirichlet}(\delta)$, the right side of equation (3.14) will be a $\text{Dirichlet}(\delta + n^{(w)}_{-i,K})$, where $n^{(w)}_{-i,K}$ is the number of assignments of the term w to the topic K, not including the $i-$th term. In equation (3.14), $\mathbf{z}_{-i}$ makes partitions to sets which are assigned in different topics. The equation (3.13) can be seen as the expectation of Dirichlet distribution; this

since $p(\phi^{(K)}|\mathrm{z}_{-i,\mathrm{w}_i}) \sim \text{Dirichlet}(\delta + n_{-i,K}^{(w)})$; so, the integral will be

$$p(w_i|z_i = K, \mathbf{z}_{-i}, \mathbf{w}_{-i}) = \frac{n_{-i,K}^{(w_i)} + \delta}{n_{-i,K}^{(\cdot)} + V\delta} \quad , \tag{3.15}$$

where $n_{-i,K}^{(\cdot)}$ is the number of times that the term w is assigned to the topic K, not including the $i-$th term.

Finding $p(z_i = K|\mathbf{z}_{-i})$ is analogous, because the expression can be seen as a multinomial distribution over topics from the document d_i, where the term w_i is modeled. Given an specific $\theta^{(d_i)}$,

$$\begin{aligned} p(z_i = K|\mathbf{z}_{-i}) &= \int p(z_i = K) p(\theta^{(d_i)}|\mathbf{z}_{-i}) d\theta^{(d_i)} \\ &= \frac{n_{-i,K}^{(d_i)} + \alpha}{n_{-i,\cdot}^{(d_i)} + k\alpha} \quad , \end{aligned} \tag{3.16}$$

where the counting of the $j-$th term in the vocabulary which is currently assigned to the topic K without the $i-$th term is $n_{-i,K}^{(j)}$.

The product between the equations (3.15) and (3.16), leads to the conditional probability

$$p(z_i = K|\mathbf{w}, \mathbf{z}_{-i}) \propto \frac{n_{-i,K}^{(w_i)} + \delta}{n_{-i,K}^{(\cdot)} + V\delta} \, \frac{n_{-i,K}^{(d_i)} + \alpha}{n_{-i,K}^{(d_i)} + k\ \alpha} \quad . \tag{3.17}$$

This result can be explained in intuitive terms; the first ratio represents the probability of a term w_i under a topic K, and the second ratio is the probability of the K topic in a document d_i. With this information, the conditional distribution can be calculated.

In this model, $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ are prior distribution parameters. So, $\boldsymbol{\delta}$ belongs to the distribution of terms over topics $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ to the distribution of documents $\boldsymbol{\theta}$. Finally, for Gibbs sampling, the

estimation equations will be given by:

$$\hat{\beta}_K^{(w_i)} = \frac{n_K^{(w_i)} + \delta}{n_K^{(\cdot)} + V\delta} \quad \text{y} \quad \hat{\theta}_K^{(d)} = \frac{n_K^{(d)} + \alpha}{n_K^{(\cdot)} + k\ \alpha} \quad , \tag{3.18}$$

for $i = 1, \ldots, V$ and $d = 1, \ldots, D$.

As it is in the presence of a mixed model, the distribution of terms given the topics, can be written as:

$$p(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(V\delta)}{\Gamma(\delta)^V}\right)^k \prod_{K=1}^{k} \frac{\prod_{j=1}^{V} \Gamma(n_K^j) + \delta}{\Gamma(n_K^{(\cdot)} + V\delta)} \quad . \tag{3.19}$$

The expressions of equation (3.19) are previously defined, and $\Gamma(\cdot)$ corresponds to the Euler gamma function. The Gibbs sampling is used to obtain the number of topics for the LDA model, given a document-term matrix.

As was discussed before, the lower bound of estimation from VEM model is used since the test likelihood is approximated using the lower bound. The log-likelihood of the Gibbs sampling will be determined by:

$$\ln p(\mathbf{w}|\mathbf{z}) = k \ln \Gamma(V\delta) - Vk \ln \Gamma(\delta) + \sum_{K=1}^{k} \left\{ \left[\sum_{j=1}^{V} \ln \Gamma(n_K^{(j)} + \delta) \right] - \ln \Gamma(n_K^{(\cdot)} + V\delta) \right\} \quad .$$

Finally, the perplexity is used as an indicator in information theory, which is a measure to examine the "state of confusion" of a model, given a number of topics. This is equivalent to the geometric mean for term of the likelihood, and it is given by:

$$\text{Perplexity}(\mathbf{w}) = \exp\left\{ -\frac{\ln p(\mathbf{w})}{\sum_{d=1}^{D} \sum_{j=1}^{V} n^{jd}} \right\} \quad , \tag{3.20}$$

where n^{jd} is the number of times that the $j-$th term occurs in the $d-$th document, and $p(\mathbf{w})$ is the topic probability, denoted as:

$$\ln p(\mathbf{w}) = \sum_{d=1}^{D} \sum_{j=1}^{V} n^{(jd)} \ln \left(\sum_{K=1}^{k} \theta_K^{(d)} \beta_K^{(j)} \right) \quad . \tag{3.21}$$

Gibbs sampling can be used to determine the weights of $\theta_K^{(d)}$. Then, the perplexity is calculated as the arithmetic mean of different models for each K simulated.

Chapter 4

Methodology for real response data

4.1 Data

Almost every semester, the Natural Science Faculty in the Universidad Nacional de Colombia runs a survey among its students. The design of the survey leads to measure professors' performance in classroom, the subject relevance of the course, and eficiency of the available resources in the campus. The measurement is run from students' perception whose responded the survey. This measurement is leaded from the students' perception. Original data include 14703 registers (surveys), which includes 606 courses, 306 subjects, and 401 professors.

For experimental purposes, a sample of 1380 surveys were kept, which includes 124 professors and 123 subjects. 36 different undergraduate programs are included in the sample. The selected subjects in the sample contain between 4 and 20 surveys, and professors have between 3 and 36 surveys answered by their students. The number of professors and subjects is not the same; this happens because there are subject with shared courses, where not all professors were scored by their corresponding students.

4.2 Categorical analysis

For categorical data, it is necessary to ensure the unidimensionality. As described before, PCA is a common used technique to assess the unidimensionality of the data. Also, the CMC was plotted, and Cronbach's alpha reliability was calculated.

After the dimensionality analysis, the professors' performance estimation was made through a Stan code, using 2000 iterations and 4 chains per run for each professor's performance and for subject relevance, deleting the first 1000 iterations as burning. Thus, 4000 samples were obtained.

After professors' performance estimation, a histogram was plotted, an increasing order subject relevance measure with a 95% credibility band, boxplots for each department of the university, and some item's cut points in Figure 5.2.

Since the parameters were simulated, these were analyzed with previously mentioned tools. The bayesian Kendall's correlation estimation was calculated for professors' performance estimation, for different combinations of professor's performance, students' severity, harmonic mean and self-performance.

A bayesian estimation of mutual information is calculated for professors' performance, students' severity, harmonic mean and self performance. For normalized mutual information, the bayesian estimator is calculated for different parameters but for but not for itself.

To explore the results of subject relevance estimation, a histogram was made to explore this measure; an increasing order by relevance measures, boxplots for each department and some cut points of item are in Figure 5.3.

For students' severity, a histogram was plotted, a dispersion diagram for subjects severity latent traits and professors' severity latent traits. Also, self performance versus professor performance and subject relevance were showed through a dispersion diagram in Figure 5.4.

Finally, a strategic diagram was plotted to show subject measure versus professor perfor-

mance. This displays four quadrants for each performance/subject combination.

4.3 Textual analysis

For the text analysis, the corpus was made only with strengths related professors' comments. This is done because when students commented about the professors' weakness, they wrote not only about the professors weaknesses, they made a catharsis. They wrote about university's infrastructure, materials, tools, etc. Given this fact, it does not make sense to model comments which are not oriented to professor weaknesses, because there are a lot of topics involved in their comments. Thus, only strength related documents where used. Remark that for textual data analysis language, each student's comment becomes a document for the corpus. The corpus was modified, removing redundant terms such as "professor", which was removed from the corpus, since it influences on the rest of the textual data analysis. The corpus was cleaned, removing punctuation and numbers from it. Each document was stemmed. This means that by each set of terms with the same morphological structure, will be kept as the same string character.

Once the text was cleaned, the corpus terms can be plotted in a word-cloud. Before applying LDA, in order to know what number of topics should be modeled, Gibbs sampling was applied, but the suggested number by the perplexity was unusable, because the document-topic distribution was equivalent to a uniform distribution. Therefore, three topics where considered for LDA modelling heuristically.

LDA was applied for the whole text corpus; giving different characteristic of professors' teaching. χ^2-distance, as it was mentioned before, is a distance measure to compare distributions, which is useful for studying distance between each topic given different documents probability. A clustering dendogram was plotted using this distance measure; next, a $K-$means clustering, to finally make applying a PCA with those centroids.

Chapter 5

Experimental results

5.1 Categorical data results

A PCA analysis was for categorical values, which can be seen in left side of the Figure 5.1. For a PCA analysis, it is (carefully) assumed that values are real. Note that the first component of PCA, explains the 41.3% of variety of the data. Thus, it is possible to conclude unidimensionality of the data.

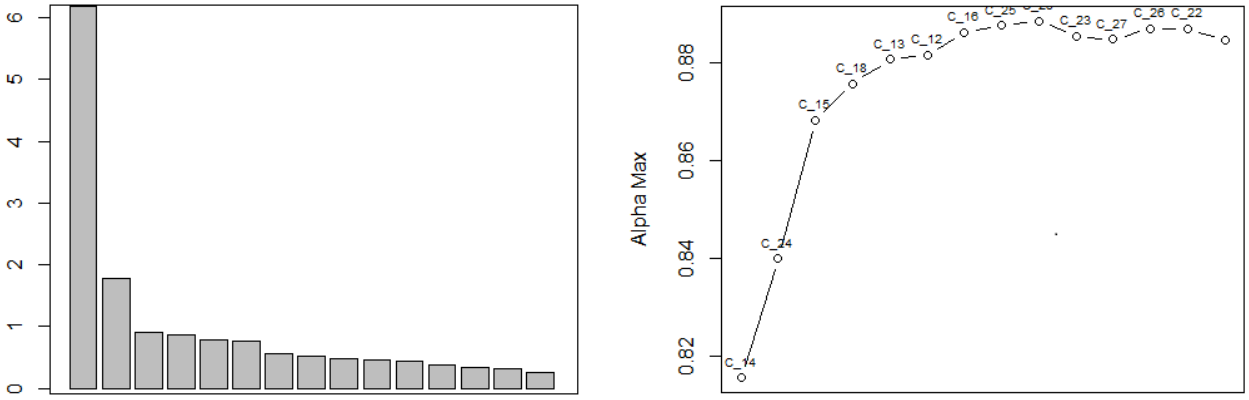

Figure 5.1: Boxplots for the eigenvalues of PCA analysis and Cronbach-Mesbah Curve

Now looking at the right side of Figure 5.1, the CMC is plotted. Item 23 (9 in appendix E) seems to deny the idea about unidimensionality, because the CMC needs to be monotonically increasing. This item can be removed, becausue it may not belong to the construct *professors' performance in the classroom*. However, it is conserved for the posterior analysis, and it is not

an heuristically notorious problem. Furthermore, the Cronbach's alpha reliability for professors' performance data was 0.892. Since the analysis of unidimensionality for subject's relevance was similar, the data is ready to run the estimation process.

5.1.1 Professors' performance estimation

The data related to professors' performance estimation consists in 15 items. The questionnaire can be found in Appendix E, which was designed to measure professors' performance. After the data management, the data set is a 20617×4 matrix. There are 124 professors, 15 items and 1380 surveys. During the preliminary analysis, it was found that the two lower categories had very few responses. Because of this, the first three categories were merged into one.

The Stan code was runned with 2000 iterations and 4 chains per run for professor performance and subject relevance measures. The burning was of 1000 iterations to obtain a total of 4000 samples from the posterior distribution in every case. This can be seen in Appendix B at line 24.

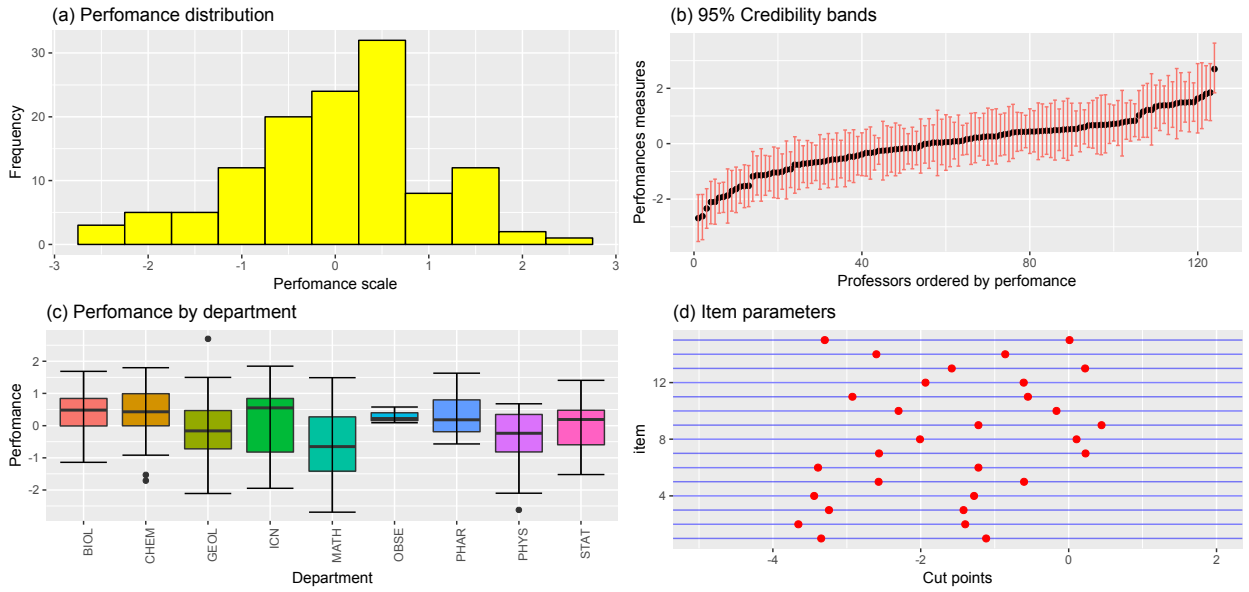

Figure 5.2: Plots of professor performance estimation. Taken from Stan output. (a) shows the histogram of the professor performance latent variable. (b) shows professor performance with credibility bands ordered by performance. (c) shows the performance by department. (d) shows cut points for some items.

The Appendix C displays the estimations computed using Stan. The $\hat{R}$ statistic for convergence diagnostic column (Gelman and Rubin, 1992) was omitted. But all values are close to one; thus, all chains are convergent. Main results of professors' performance estimation can be seen in Figure 5.2 (b).

The performance distribution at Figure 5.2 (a) shows the estimate latent trait. It was supposed to follow a standard normal distribution; that is, $N(0, 1)$. The plot at the right side of it, shows the performance estimation for each professor, ordered in increasing order by it estimation; with 95% credibility bands. The (c) plot is the distribution of the latent trait by departments. The comparison of the performance is done since the performance estimation values are a real numbers. A general image of the test can be seen at (c). In this plot, each line represents an item,

and colored dots are the cut points for each item. The interpretation of the item can be read as follows: the item 2 is the *easiest* item, in a generic sense; this is because students tend to consider that professors are good in this regard. The item 2 in the questionnaire is the question "Does the professor respect the dates agreed for academic activities, including evaluations and delivery of results?". In contrary case, the *hardest* item will be the 9^{th}, in a general sense students think professors are not very good at this. In this case, the question was "Does the professor organize activities that allow me to exercise my oral and written capabilities?". Note that this item may not belong to the same construct as *professors' performance in the classroom*, because in the CMC, this item has a decreasing curve.

It can be seen in Appendix C that the generated quantities μ_β are useful to do these analyzes. Note that extreme values of μ are $\mu_{\beta_2} = -2,53$ and $\mu_{\beta_9} = -0.39$. Remember that $\mu_{\beta_k}-$values are estimation means of $\mu_{\beta_{kj}}$ cutting points. Moreover, these values have small variance because the β_{jk}'s distributions is generated given that the samples for each item is large.

(params, stats)	Q2.5	Q25	Q50	mean	Q75	Q97.5
(θ, γ)	-0.050	-0.027	-0.015	-0.014	-0.002	0.022
(θ, ρ)	0.743	0.774	0.791	0.791	0.808	0.839
(θ, ξ)	0.142	0.160	0.169	0.169	0.178	0.195
(γ, ρ)	-0.119	-0.095	-0.082	-0.081	-0.068	-0.043
(γ, ξ)	-0.206	-0.190	-0.182	-0.182	-0.173	-0.157

Table 5.1: Bayesian estimation of Kendall correlation estimation in professor performance estimation. Where θ is professor performance, γ is student severity, ρ is the harmonic mean, and ξ is self-performance.

Tables 5.1, 5.2, and 5.3 show the bayesian estimation of Kendall correlation, mutual information and normalized mutual information respectively, for latent variables of interest, described as in methodology chapter. For comparisons purposes, the harmonic mean of professors' scores are included. As was expected, students' severity and professors' performance are independent latent variables. On the another hand, a small positive association between self-performance and

params/stats	q2.5	q25	q50	mean	q75	q97.5
(θ, θ)	1.600	1.605	1.607	1.607	1.608	1.610
(θ, γ)	-0.005	-0.001	0.001	0.001	0.003	0.009
(θ, ρ)	0.602	0.668	0.707	0.709	0.747	0.834
(θ, ξ)	0.019	0.023	0.026	0.026	0.028	0.033
(γ, γ)	2.401	2.401	2.401	2.401	2.401	2.401
(γ, ρ)	-0.001	0.003	0.006	0.006	0.009	0.015
(γ, ξ)	0.021	0.026	0.029	0.029	0.032	0.039
(ρ, ρ)	1.609	1.609	1.609	1.609	1.609	1.609
(ρ, ξ)	0.027	0.027	0.027	0.027	0.027	0.027
(ξ, ξ)	0.969	0.969	0.969	0.969	0.969	0.969

Table 5.2: Bayesian estimation of mutual information in professor performance estimation. Where θ is professor performance, γ is student severity, ρ is the harmonic mean, and ξ is self-performance.

professors' performance is found, and a small negative association between self-performance and students' severity.

5.1.2 Subject relevance estimation

The matrix data has a dimension of 5492×4 for this case. After merge categories 1 and 2, there are only 4 categories for each item; this procedure was replicated for four items. Also, there are 1380 surveys and 123 subjects. In Figure 5.3, main results are showed for subject relevance estimation, Appendix D include the estimated parameters. The instrument designed to measure subject relevance can be found in Appendix F. Note that $\sigma_\beta = 2.41$ is almost twice than the corresponding value in the professor performance estimation. This is an implication of the number of items used for this case. Nevertheless, since these items have more categories, they have more precision. But then, the standard deviation of student severity is $\sigma_\gamma = 1.27$, which is similar to the same parameter for professor performance estimation.

(params/stats)	Q2.5	Q25	Q50	mean	Q75	Q97.5
(θ, γ)	-0.003	-0.001	0.001	0.001	0.002	0.004
(θ, ρ)	0.374	0.416	0.440	0.441	0.465	0.518
(θ, ξ)	0.015	0.019	0.020	0.021	0.022	0.027
(γ, ρ)	-0.001	0.002	0.003	0.003	0.005	0.008
(γ, ξ)	0.014	0.017	0.019	0.019	0.021	0.025
(ρ, ξ)	0.021	0.021	0.021	0.021	0.021	0.021

Table 5.3: Bayesian estimation of normalized mutual information in the professor's performance estimation. Where θ is professor performance, γ is student severity, ρ is the harmonic mean, and ξ is self-performance.

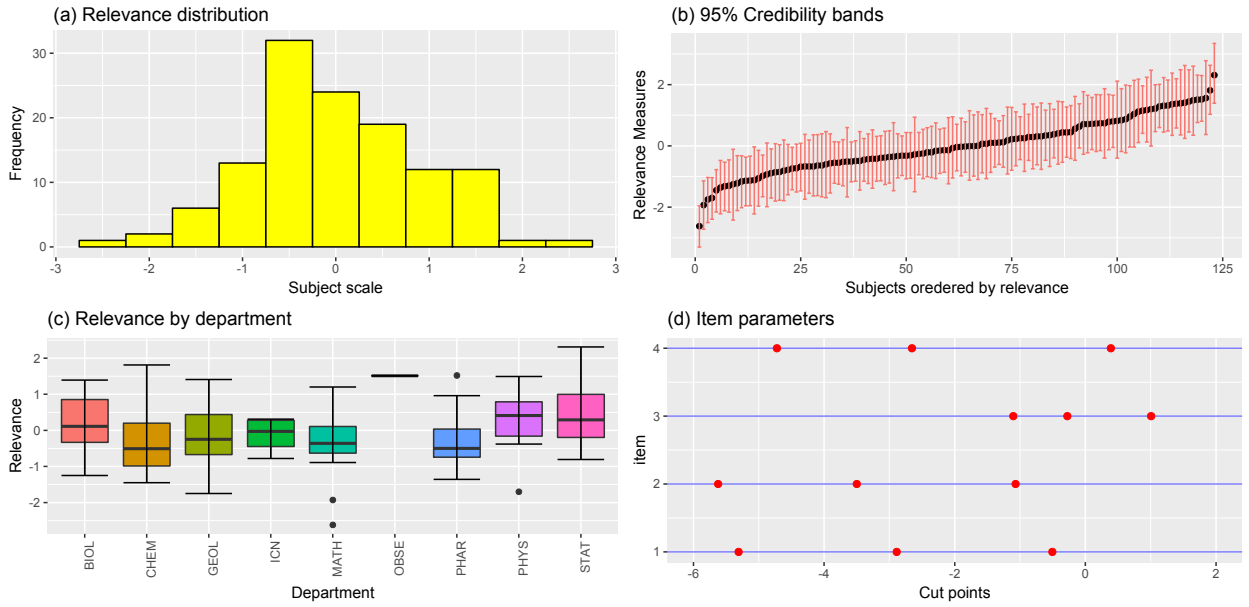

Figure 5.3: Plots from subject relevance estimation. Computed from the Stan output.

Compared to the previous results, the interpretations above are analogue. However, for this case, the interpretation must be regarding to latent estimated parameters. These results correspond to the perception of the students about their subjects, which they are currently coursing. For evaluation of academic programs, this is a useful information. For instance, the item 3 cuts indicate that students consider such subject is not useful for their professional future. This is true because the question was "Do you consider that the theoretical and practical contents of the subject are useful for the professional activity?". This bad result for the subject may be associated with a list

of factors which influences their answers, such as their experience, particular interests, and so on. An important feature of Natural Science Faculty of Universidad Nacional de Colombia is that offers many theoretical courses for the most of its undergraduate careers.

5.1.3 Severity analysis

To analyze the survey, is very important the inclusion of severity/clemency parameters. This will correct problems associated with the use of the survey, in order to measure the professor performance.

The severity estimation information is plotted in Figure 5.4. The top left graph shows the histogram of the distribution of estimated severity of 1380 students, computed from the professor's performance estimation. At the right side of it, show the relationship between the professor performance measurement and the subject relevance measurement.

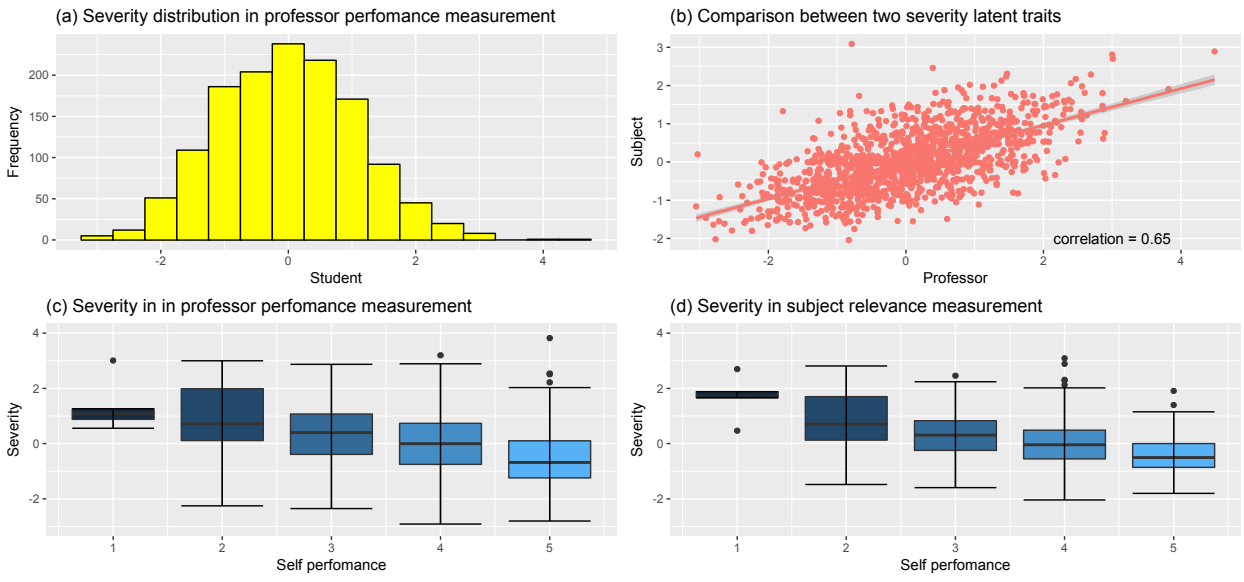

Figure 5.4: Severity analysis.

All students who responded the survey are anonymous. As was mentioned before, experts recommend to include some socio-demographic questions related to students' information, which could influence their perception or point of view about the professor performance. Also, students' information such as enrollments, number of times the student have disapproved the subject, and a self-evaluation of their performance in the course. Self-evaluation and severity has an inverse relationship. For this case, the Likert scale are integer values from 1 to 5. This is a good approximation about the expected score by student in the course. The distribution of the self-performance variable is shown in Table 5.4. As was mentioned before, lower scale values have lower frequency.

Value	Frequency
1	5
2	22
3	246
4	890
5	217

Table 5.4: Distribution of the self-performance variable in the sample.

A common sense expectation, is that self-evaluation and student severity have an inverse relationship. That is, the more severe a student is, the lower is the self-evaluation score. Nevertheless, this association between these variables is small in the data.

5.1.4 Strategic diagram

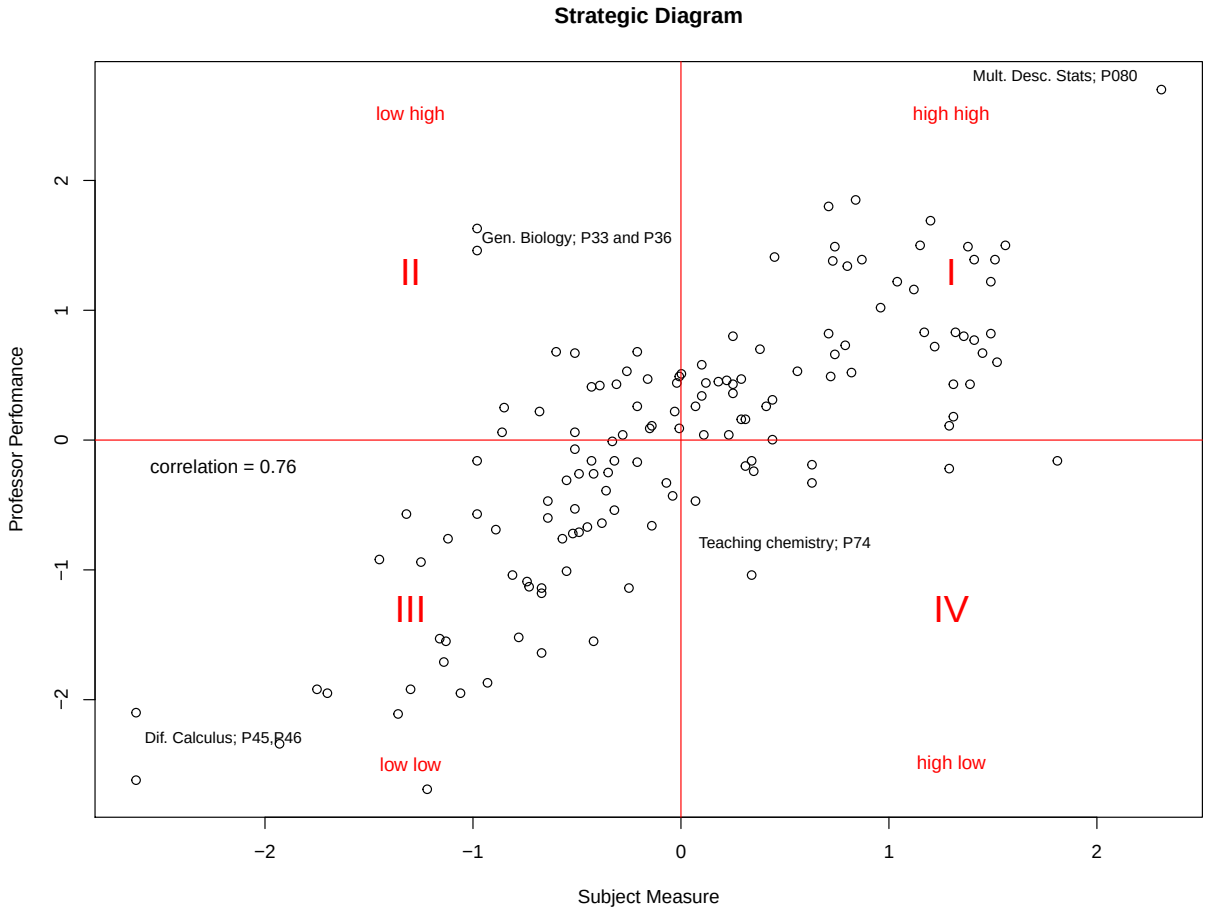

Figure 5.5: Strategic diagram.

An strategic diagram is showed in Figure **??**. This graph displays different areas which can help to make decisions. In the diagram, latent trait variables are crossed; these variables are professor performance and subject relevance. Professors and subjects are organized in an interesting way. In the first quadrant, at top right of Figure **??**, are located those professors with a high performance, whose classes are considered the most relevant subjects. In quadrant II, at the left size of the first one, are located with low evaluated relevance subject for each career, but which professors

are good evaluated by their students. The third quadrant has low subject relevance from students' perspective, with professors with a low teaching performance. Finally, at fourth quadrant, higher relevance subject are located, but with high performance teachers.

The strategic diagram can lead to a variety of interpretations. Also, it can be used to select professors for the most relevant subjects in the career.

5.2 Textual data analysis results

For textual data analysis, the first approach is to plot frequencies of words in the corpus. Figure 5.6 shows a barplot graph for the term frequencies of those terms with frequency bigger than 20. Another common used tools for visualizing textual data is the word-cloud.

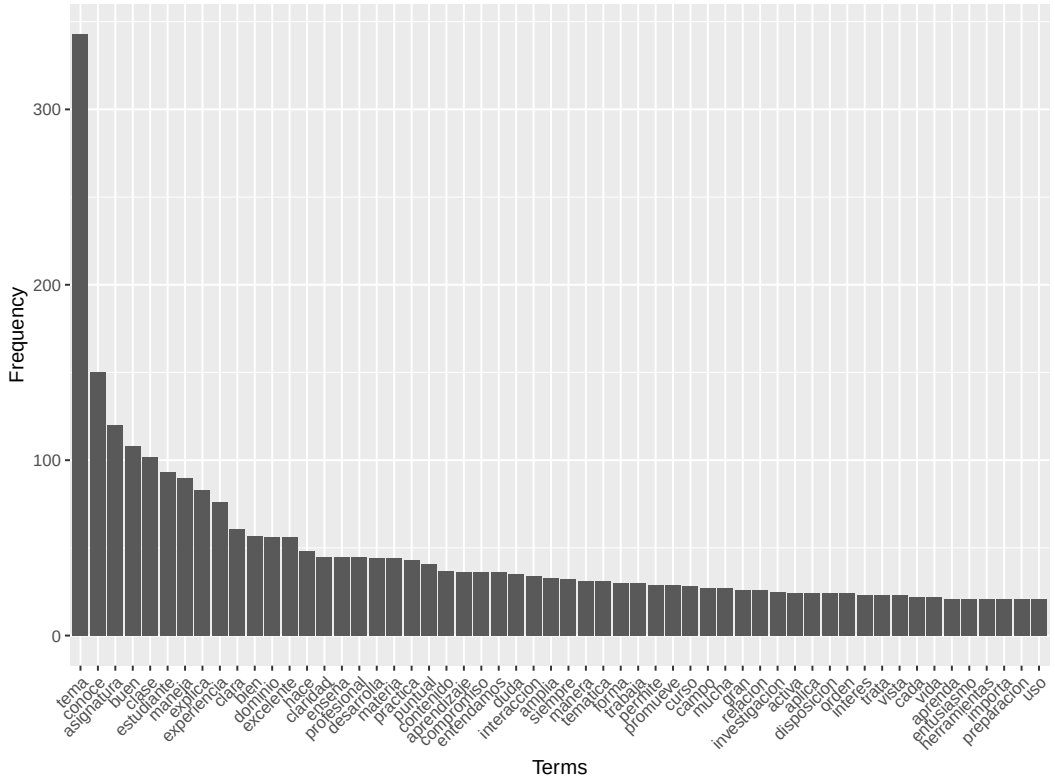

Figure 5.6: Term frequencies of the corpus.

The word-cloud in Figure 5.7 reveals that the term *tema*, which means topic, is the most common used word among the comments. Other terms like *conoce* (knows), *asignatura* (subject), *clase* (class), *bueno* (good), *estudiante* (student), *maneja* (proficiency).

Figure 5.7: Word-cloud with frequency bigger than 20.

Before applying LDA with certain number of topics, in order to calculate the perplexity with estimated parameters for each k, the Gibbs sampling was simulated for $k = 1, \ldots, 30$ The Gibbs sampling was runned with next described simulation parameters. The burning was of 1000, which allow unbias chains. The simulations were kept every 100 iterations. A process called "thinning". In total, 10000 iterations were saved for each k. The estimated perplexity for each k is plotted in Figure 5.8.

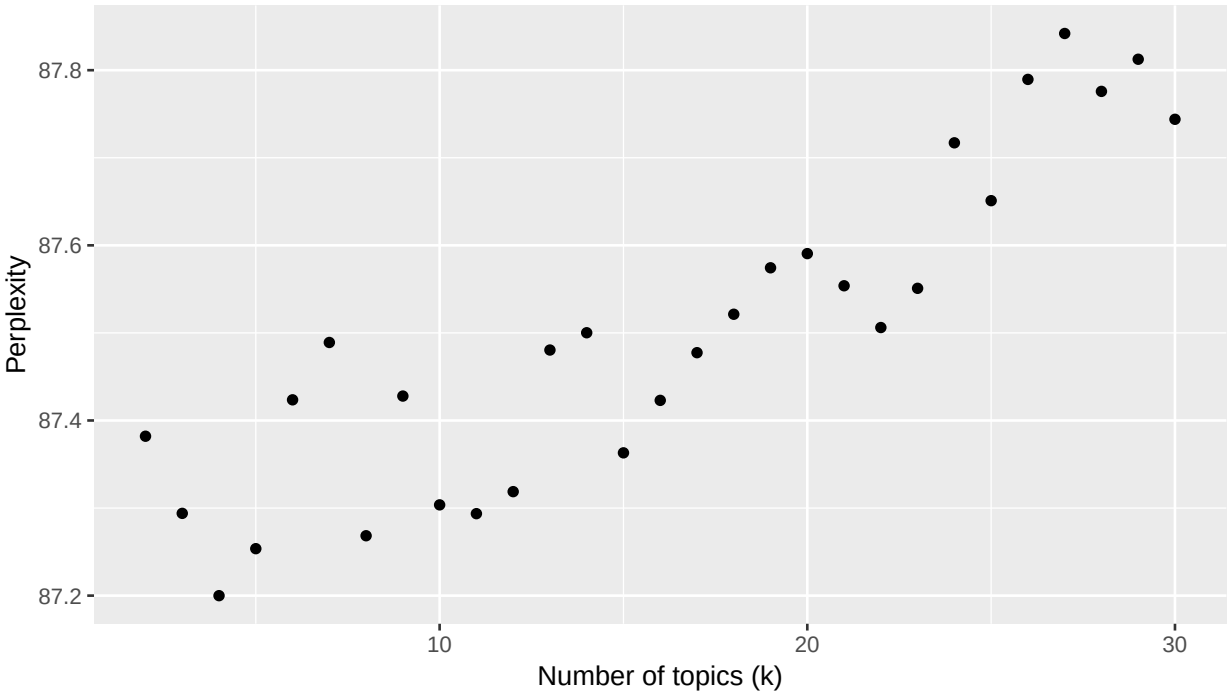

Figure 5.8: Perplexity using Gibbs sampling for $k = 2, \ldots, 30$, applied for the corpus.

First, the LDA model was fitted with 3 topics. The issue with this number of topics was that document given topic distribution was heuristically equivalent to the uniform distribution, because the probability for each document given a topic was similar for each topic. But, when fitting the LDA model using 4 topics, those probabilities are not the same for each document given document.

Document	Topic_1	Topic_2	Topic_3
1	0.41	0.31	0.28
2	0.28	0.30	0.42
3	0.30	0.34	0.36
4	0.40	0.32	0.28
5	0.38	0.31	0.31
6	0.39	0.31	0.30

Table 5.5: First six values of estimated probabilities of topics given documents.

The Figure 5.9 shows the top 10 terms with higher term given topic probability for each topic. On one side, a topic given document distribution is plotted by colors.

Topic Distribution

Knowledge	Teaching	Development
knows	class	theme
subject	explain	good
student	clear	manage
develop	good	experience
matter	does	domain
content	excellent	puntual
compromise	teach	practice
interact	clarity	learn
tematic	always	doubt
work	way	large

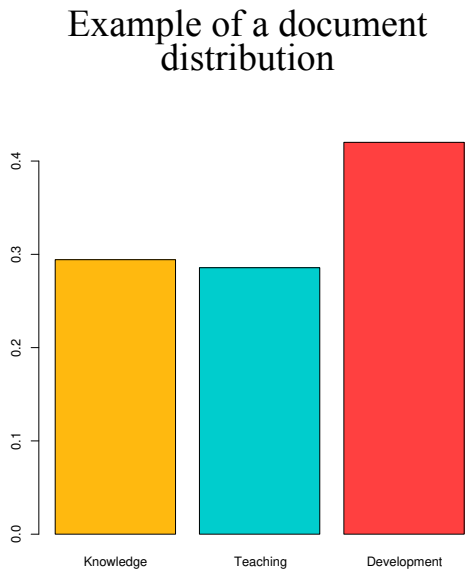

Figure 5.9: Topic distribution for different terms and documents.

Note that in Figure 5.9 the topics are named by the terms they contain. The name of each topic is chosen heuristically.Then, χ^2-distance is calculated for each pair of documents, as in

equation (2.13). Using this equation, a distance matrix can be constructed, which is used to plot a dendogram in Figure 5.10.

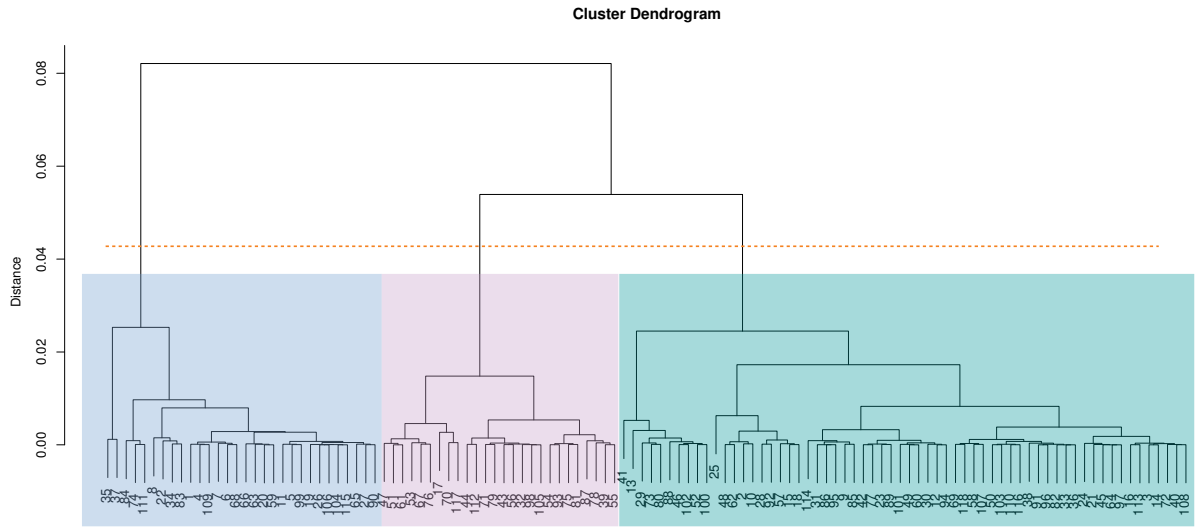

Figure 5.10: Dendogram for topics given the probability distribution of documents.

This is a useful distance, as was explained before, since it is a widely used tool to compare distributions. The next step is to use $K-$means clustering, in order to see and explore possible centroids of interest. Before to apply $K-$means, a reasonable number of centroids must be given. For this reason, note that there are three topics, which indicate professors' strength. Given this, it is reasonable to think that professors can have different combinations of this three strength aspects at the barplot (as it is shown at right side in Figure 5.9), or they have an equal probability of topics given documents. This thinking leads to include 4 centroids for a clustering analysis. The Table 5.6 shows the $K-$means results for $K = 4$ and 5000 iterations.

Centroid	Topic_1	Topic_2	Topic_3
1	0.29	0.30	0.41
2	0.33	0.32	0.35
3	0.28	0.40	0.32
4	0.42	0.30	0.29

Table 5.6: $K-$means centroids using χ^2-distance, with $K = 4$.

Finally, a PCA is applied already knowing that topics given documents distribution can be clustered using 4 centroids. So, the PCA is applied with the indexation of 4 centroids for each topic given document probability. The Figure 5.11 shows the different professors' strengths, where each centroid was labeled using different strengths topics. This plot shows different abilities of professor's measured only through students' comments, and it could be useful to measure the professors' strengths.

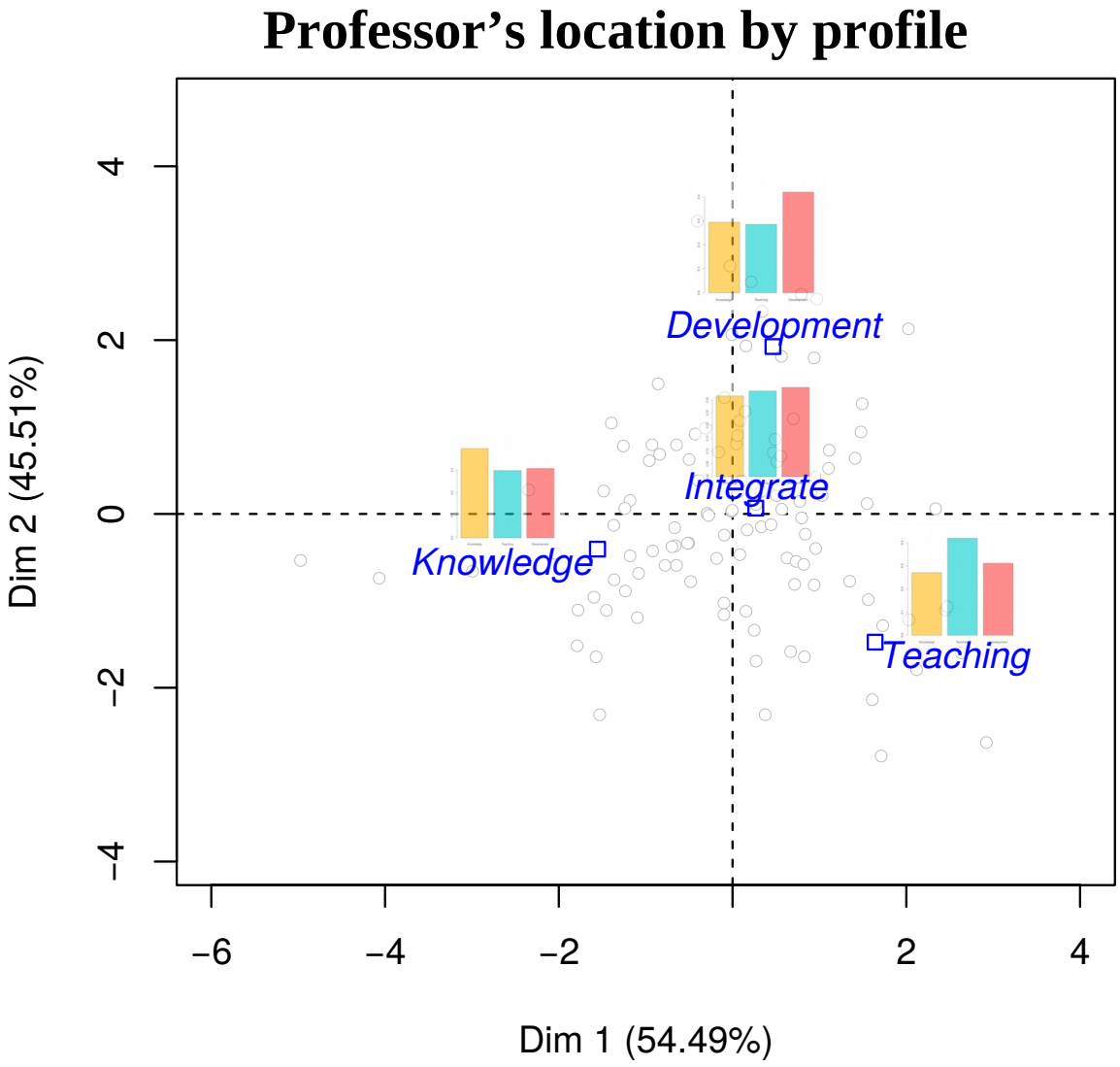

Figure 5.11: PCA plot using 4 indexed centroids from clustering.

Chapter 6

Conclusions and discussions

The use of multi-faceted Rasch models and latent Dirichlet Allocation applied to educational data surveys provide an improved SET analysis. This can help to academic units to improve the performance of their courses through the right choice of professors for very relevant courses. An important issue related to this, is that students give more importance to core courses in their programs. The analysis used in this work propose an integral analysis for courses, studying the student responses through different items, and textual comments.

From the categorical analysis, IRT models were used in order to estimate different professors' an students' characteristics. Also, an strategic diagram is provided to study different subjects and professors by their performance and relevance. The strategic diagram can lead to a variety of interpretations. Also, it can be used to select professors for the most relevant subjects in the career. It also displays those subjects that need to be improved putting another professor in that position.

From the textual analysis results, a natural conclusion is that the survey can be improved through asking for subject and faculty strengths and weaknesses. This should work since students answered about these topics when they were asked for professors' weaknesses. Using a more polished survey should improve the results, even for textual comments.

An important result from the textual analysis, is the topic distribution. This explains what professor strengths are students commenting about. Since these topics, clustering methods allow

to measure the centroids with χ^2-distance.

A strategic diagram is proposed, through principal component analysis. PCA clusters professors by their strengths commented by students. Weaknesses were not analyzed since student made a catharsis about to faculty deficiencies. That is, they mentioned infrastructure deterioration, old laboratory machines, and so on.

PCA is provided in the textual analysis, but when looking at "integrate" centroid, it may seem suspicious from the next point of view. Given that professors' ability barplots have the same height, intuitively, there are two possibilities: the first one is that professors with the same profile have all strengths with the same level; and the other one is the contrary case; that is, professors with such profiling do not have any of those strengths. But, given that only were analyzed strengths related comments, it is assumed that it is not possible to be in the presence of the lack of those qualities.

Bibliography

Aigner, D. and Thum, F. (1986). On student evaluations of teaching ability. Journal of Economic Education, 17 (Fall):243–266.

Apodaca, P. and Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni-and multidimensional models. Studies in Higher Education, 30(6):723–748.

Ash, R. (1990). Information Theory. Dover.

Baker, F. B. and Seok-Ho, K. (2004). Item Response Theory. Marcel Decker Inc., 2nd edition.

Bartholomew, D. and Knott, M. (1999). Latent variable models and factor. Analysis, 2nd Ed., Arnold, London.

Bartholomew, D., Knott, M., and Moustaki, I. (2011). Latent Variable Models and Factor Analysis. A Unified Approach. Wiley, third edition.

Becker, W. and Watts, M. (1999). How departments of economics should evaluate teaching. american economic review. American Economic Review (Papers and Proceedings), 89:344–349.

Birnbaum, A. (1968). Statistical Theories of mental test Scores, chapter Trait models and their use in infering an examinee's ability. Reading, MA: Addison Wesley.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

Bock, R. D. (1997). A brief history of item response theory. Educational Measurement: Issues and Practice, 16(4):21–32.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. Psychometrika, 46(4):443–459.

Box, G. (1980). Sampling and bayes' inference in scientific modelling and robustness. Journal of the Royal Statistical Society. Series A., 143:383–430.

Bucker, M., Corliss, G., Hovland, P., Naumann, U., and Norris, B., editors (2006). Automatic Differentiation: Applications, Theory, and Implementations. Springer.

Cameletti, M. and Caviezel, V. (2015). Package CMC. CRAN R-project.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. Journal of Statistical Software, Articles, 76(1):1–32.

Cohen, J. (1983). The cost of dichotomization. Applied psychological measurement, 7(3):249–253.

Cover, T. M. and Thomas, J. A. (1991). Elements of Information Theory.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16:297 – 334.

Daliri, M. R. (2013). Chi-square distance kernel of the gaits for the diagnosis of parkinson's disease. Biomedical Signal Processing and Control, 8(1):66–70.

Decanio, S. (1986). Student evaluations of teaching — a multinominal logit approach. Journal of Economic Education, 17 (Summer):165–176.

Dilts, D. (1980). A statistical interpretation of student evaluation feedback. Journal of Economic Education, 14 (Spring):1–5.

Eckes, T. (2015). Introduction to Many-Facet Rash Measurement. Analyzing and Evaluating Rater-Mediated Assesments. Peter Lang Edition, second edition.

Fox, J.-P. (2010). Bayesian item response modeling: Theory and applications. Springer Science & Business Media.

Gelfand, A. E., Dey, D., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. Technical report, DTIC Document.

Gelman, A., Meng, X.-L., and Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica sinica, 6(4):733–760.

Gelman, A. and Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. Statist. Sci., 7:457–472.

Greenacre, M. (2017). Correspondence analysis in practice. Chapman and Hall/CRC.

Hambleton, R. K. (1991). Fundamentals of Item Response Theory (Measurement Methods for the Social Science). Sage, 1 edition.

Hoffman, M. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. Journal of Machine Learning Research, 15.

Holland, P. W. (1990). On the sampling theory roundations of item response theory models. Psychometrika, 55(4):577–601.

Jollife, I. (2002). Principal Component Analysis. Springer, 2nd edition.

Kavalseth, T. O. (2017). On normalized mutual information: Measure derivations and properties. Entropy, 19:631 – 644.

Krautmann, A. C. and Sander, W. (1999). Grades and student evaluations of teacher. Economics of Education Review, 18:59–63.

Lehmann, E. L. and Casella, G. (2006). Theory of point estimation. Springer Science & Business Media.

Lord, F. and Novick, M. (2013). Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company.

Lord, F. M. and Novick, M. R. (2008). Statistical theories of mental test scores. IAP.

Luo, Y. and Jiao, H. (2017). Using the stan program for bayesian item response theory. Educational and Psychological Measurement, 2017:1–25.

Masters, G. N. (1982). A rasch model for partial credit scoring. Psychometrika, 47(2):149–174.

Mehdizadeh, M. (1990). Loglinear models and student course evaluations. Journal of Economic Education, 21 (Winter):7–21.

Meyer, P. E. (2015). Package Infotheo. CRAN R-project.

Molenaar, I. W. (1995). Some background for item response theory and the rasch model. In Rasch Models, pages 3–14. Springer.

Muraki, E. (1992). A generalized partial credit model: Application of an em algorithm. ETS Research Report Series, 1992(1):i–30.

Muraki, E. (1993). Information functions of the generalized partial credit model. ETS Research Report Series, 1993(1):i–12.

Neal, R. (2011). MCMC using Hamiltonian dynamics in Handbook of Markov Chain Monte Carlo, pages 113–162. New York, NY: CRC Press.

Rabe-Hesketh, S. and Skrondal, A. (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Chapman and Hall/CRC.

Rasch, G. (1960). Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests.

Robert, C. and Casella, G. (2010). Introducing Monte Carlo Methods with R. Springer Science & Business Media.

RStan, D. T. (2017). RStan: the R interface to Stan. R package version 2.16.2.

Samejima, F. (1997). Graded response models. In Handbook of Item Response Theory, Volume One, pages 85–100. Chapman and Hall/CRC.

Seiver, D. (1983). Evaluations and grades: a simultaneous framework. Journal of Economic Education, 14 (Summer):32–38.

Spearman, C. (1904). The proof and measurement of association between two things. American journal of Psychology, 15:72–101.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Gilks, W. (1997). BUGS: Bayesian inference using Gibbs sampling, Version 0.60. Cambridge: Medical Research Council Biostatistic Unit.

Stark, P. and Freishtat, R. (2014). An evaluation of course evaluation. ScienceOpen Research-Section SOR-EDU, pages 1–7.

Thissen, D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47(2):175–186.

Uttl, B., White, C., and Wong, D. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. Studies in Educational Evaluation, 54:22–42.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. Assessment and Evaluation in Higher Education, 23(2):191.

Appendix A

Stan code for estimating professor performance

```
01

02  data{
03  int<lower=1>  N;                          // number of responses
04  int<lower=10> N_prof;                     // number of professors
05  int<lower = 10> N_stud;                   // number of students
06  int<lower=2>  N_item;                     // number of items
07  int<lower=2,upper=5> N_cat;               // number of categories
08  int<lower=1,upper=N_cat> y[N];            //  y[n], n-th response
09  int<lower=1,upper=N_prof> professor[N]; // professor of response[n]
10  int<lower=1,upper=N_item> item[N];      // item of response[n]
11  int<lower=1,upper=N_stud> student[N];   // professor of response[n]
12  }
13

14 parameters{
15  vector[N_prof] theta;          // latent traits of professors
16  vector[N_stud] gamma;          // students'severity
17  ordered[N_cat-1] beta[N_item]; // beta parameters
18  real mu_beta;                  // mean of the beta-parameters
19  real<lower=0> sigma_beta;      // sd of beta prior distributions
```

```
20  real<lower=0> sigma_gamma;     // sd of gamma prior distributions
21  }
22
23 model{
24  theta ~ normal(0,1);              // scale of the latent traits
25  gamma ~ normal(0,sigma            //scale of the students'severity
26  for(j in 1:N_item){
27   beta[j]~ normal(mu_beta,sigma_beta); // prior of betas
28   }
29  mu_beta ~ normal(0,2);     // hyper prior for mu_beta
30  sigma_beta ~ cauchy(0,2);  // hyper prior for sigma_beta
31  sigma_gamma ~ cauchy(0,2); // hyper prior for sigma_gamma
32  //likelihood
33  for(n in 1:N){
34   y[n]~ ordered_logistic(theta[professor[n]]-gamma[student[n]],beta[item[n]]);
35   }
36 }
37
38 generated quantities {
39  vector[N_item] mu_item_beta; // the mean of betas by item
40  for (n in 1: N_item){
41    mu_item_beta[n] = mean(beta[n]);
42   }
```

Appendix B

RStan code for estimating professor performance

```
01
02 library(rstan)
03 rstan_options(auto_write = TRUE)
04 options(mc.cores = parallel::detectCores())
05
06 # Read the data
07 setwd("D:/Alvaro/Alvaro_2018/Proyecto_Edificando")
08 load(file="professor_data_sample.Rdata")
09
10 # total parameters
11 NN = nrow(professor_data) #20617; num. registers
12 NS = length(unique(professor_data[,2])) #1380; num. students
13 NP = length(unique(professor_data[,3])) #124;  num. professors
14 NI = length(unique(professor_data[,4])) #15;   num. items
15 NC = max(professor_data[,1]) #3; number of categories
16 #
17 # go to Stan
18 dat = list(professor=professor_data[,3], student = professor_data[,2],
   item = professor_data[,4], y = professor_data[,1], N=NN, N_item=NI, N_prof=NP,
```

```
   N_cat=NC, N_stud=NS)
19
20 #first compile and save the fitted model for re-using
21 # data are taken from the current R works
22 prof_fit_p_ <- stan(file = 'Multi_Faceted_Ciencias_AM_002.stan', data = dat,
   iter = 4, chains = 1)
23 # now sample using the compiled model
24 prof_fit_2<- stan(fit = prof_fit_p_,  data =dat, iter = 2000, chains = 4,
   control = list(max_treedepth = 12))
25 # save the stan object
26 save(prof_fit_2,file="Model_2_prof_fit_2_pr_it_st.Rdata")
```

Appendix C

Estimated item parameters in professor performance measurement

parameter	mean	se mean	sd	q2.5	q25	q50	q75	q97.5	n eff
$\beta_{1,1}$	-3.35	0.00	0.15	-3.63	-3.45	-3.34	-3.25	-3.06	874.02
$\beta_{1,2}$	-1.12	0.00	0.12	-1.36	-1.20	-1.12	-1.04	-0.88	690.10
$\beta_{2,1}$	-3.66	0.00	0.16	-3.97	-3.76	-3.65	-3.55	-3.36	1004.33
$\beta_{2,2}$	-1.40	0.00	0.12	-1.64	-1.48	-1.40	-1.32	-1.18	685.69
$\beta_{3,1}$	-3.24	0.00	0.14	-3.52	-3.34	-3.24	-3.15	-2.97	850.75
$\beta_{3,2}$	-1.42	0.00	0.12	-1.67	-1.50	-1.42	-1.34	-1.19	655.63
$\beta_{4,1}$	-3.44	0.00	0.15	-3.74	-3.54	-3.44	-3.34	-3.15	945.59
$\beta_{4,2}$	-1.28	0.00	0.12	-1.52	-1.36	-1.28	-1.19	-1.04	774.35
$\beta_{5,1}$	-2.57	0.00	0.13	-2.83	-2.66	-2.57	-2.48	-2.32	796.05
$\beta_{5,2}$	-0.60	0.00	0.12	-0.84	-0.68	-0.60	-0.52	-0.38	674.06
$\beta_{6,1}$	-3.39	0.00	0.14	-3.69	-3.49	-3.39	-3.29	-3.10	845.08
$\beta_{6,2}$	-1.22	0.00	0.12	-1.46	-1.30	-1.22	-1.14	-0.99	701.98
$\beta_{7,1}$	-2.56	0.00	0.13	-2.83	-2.65	-2.56	-2.47	-2.32	768.23
$\beta_{7,2}$	0.23	0.00	0.12	-0.00	0.15	0.23	0.31	0.45	682.04
$\beta_{8,1}$	-2.01	0.00	0.13	-2.26	-2.09	-2.00	-1.92	-1.77	822.66
$\beta_{8,2}$	0.11	0.00	0.12	-0.12	0.02	0.11	0.19	0.33	755.45
$\beta_{9,1}$	-1.22	0.00	0.12	-1.46	-1.30	-1.22	-1.14	-0.99	692.15
$\beta_{9,2}$	0.44	0.00	0.12	0.21	0.37	0.45	0.53	0.66	682.42
$\beta_{10,1}$	-2.30	0.00	0.13	-2.56	-2.39	-2.30	-2.21	-2.05	780.40
$\beta_{10,2}$	-0.16	0.00	0.12	-0.40	-0.24	-0.17	-0.08	0.06	700.51
$\beta_{11,1}$	-2.92	0.00	0.14	-3.19	-3.02	-2.92	-2.83	-2.64	866.12
$\beta_{11,2}$	-0.55	0.00	0.12	-0.78	-0.63	-0.55	-0.47	-0.32	716.25
$\beta_{12,1}$	-1.94	0.00	0.12	-2.18	-2.02	-1.94	-1.85	-1.70	735.21
$\beta_{12,2}$	-0.61	0.00	0.12	-0.84	-0.68	-0.61	-0.53	-0.38	716.44
$\beta_{13,1}$	-1.58	0.00	0.12	-1.82	-1.66	-1.58	-1.50	-1.34	703.54
$\beta_{13,2}$	0.22	0.00	0.12	-0.01	0.15	0.22	0.30	0.45	724.76
$\beta_{14,1}$	-2.60	0.00	0.13	-2.86	-2.69	-2.60	-2.51	-2.35	766.61
$\beta_{14,2}$	-0.86	0.00	0.12	-1.09	-0.94	-0.86	-0.78	-0.63	692.49
$\beta_{15,1}$	-3.30	0.00	0.14	-3.58	-3.40	-3.30	-3.20	-3.02	864.62
$\beta_{15,2}$	0.01	0.00	0.12	-0.22	-0.07	0.01	0.09	0.24	686.11
μ_β	-1.58	0.00	0.26	-2.10	-1.76	-1.59	-1.41	-1.07	4000.00
σ_β	1.33	0.00	0.18	1.02	1.20	1.31	1.44	1.74	4000.00
σ_γ	1.27	0.00	0.04	1.20	1.25	1.27	1.29	1.34	2241.92
μ_{β_1}	-2.23	0.00	0.12	-2.47	-2.32	-2.23	-2.15	-1.99	681.98
μ_{β_2}	-2.53	0.00	0.13	-2.78	-2.61	-2.53	-2.44	-2.29	717.43
μ_{β_3}	-2.33	0.00	0.12	-2.58	-2.41	-2.33	-2.25	-2.10	660.83
μ_{β_4}	-2.36	0.00	0.12	-2.61	-2.44	-2.36	-2.28	-2.12	737.95
μ_{β_5}	-1.59	0.00	0.12	-1.82	-1.67	-1.59	-1.51	-1.37	647.99
μ_{β_6}	-2.31	0.00	0.12	-2.55	-2.38	-2.30	-2.22	-2.07	665.69
μ_{β_7}	-1.17	0.00	0.11	-1.40	-1.25	-1.17	-1.09	-0.95	625.85
μ_{β_8}	-0.95	0.00	0.12	-1.18	-1.03	-0.95	-0.87	-0.74	695.23
μ_{β_9}	-0.39	0.00	0.11	-0.61	-0.46	-0.39	-0.31	-0.17	627.42
$\mu_{\beta_{10}}$	-1.23	0.00	0.12	-1.46	-1.31	-1.23	-1.16	-1.01	655.83
$\mu_{\beta_{11}}$	-1.74	0.00	0.12	-1.97	-1.82	-1.74	-1.66	-1.51	674.22
$\mu_{\beta_{12}}$	-1.27	0.00	0.12	-1.51	-1.35	-1.27	-1.20	-1.05	668.39
$\mu_{\beta_{13}}$	-0.68	0.00	0.11	-0.91	-0.75	-0.68	-0.60	-0.46	647.90
$\mu_{\beta_{14}}$	-1.73	0.00	0.12	-1.97	-1.81	-1.73	-1.65	-1.50	649.50
$\mu_{\beta_{15}}$	-1.64	0.00	0.12	-1.87	-1.72	-1.64	-1.56	-1.42	647.08

Appendix D

Estimated item parameters in subject relevance measurement

parameter	mean	se mean	sd	q2.5	q25	q50	q75	q97.5	n eff
$\beta_{1,1}$	-5.31	0.00	0.26	-5.82	-5.47	-5.30	-5.14	-4.82	2228.27
$\beta_{1,2}$	-2.89	0.00	0.14	-3.18	-2.99	-2.88	-2.79	-2.60	1504.25
$\beta_{1,3}$	-0.51	0.00	0.12	-0.74	-0.59	-0.51	-0.42	-0.27	1355.19
$\beta_{2,1}$	-5.62	0.01	0.29	-6.21	-5.80	-5.62	-5.43	-5.07	2589.55
$\beta_{2,2}$	-3.50	0.00	0.16	-3.82	-3.60	-3.50	-3.40	-3.19	1649.97
$\beta_{2,3}$	-1.07	0.00	0.12	-1.31	-1.15	-1.07	-0.98	-0.83	1331.75
$\beta_{3,1}$	-1.10	0.00	0.12	-1.34	-1.19	-1.11	-1.02	-0.86	1258.74
$\beta_{3,2}$	-0.28	0.00	0.12	-0.51	-0.35	-0.28	-0.20	-0.04	1332.86
$\beta_{3,3}$	1.01	0.00	0.12	0.77	0.92	1.00	1.09	1.25	1447.58
$\beta_{4,1}$	-4.72	0.00	0.21	-5.15	-4.86	-4.72	-4.58	-4.32	2005.34
$\beta_{4,2}$	-2.66	0.00	0.14	-2.92	-2.75	-2.65	-2.56	-2.39	1459.10
$\beta_{4,3}$	0.39	0.00	0.12	0.16	0.31	0.39	0.47	0.62	1440.03
μ_β	-1.94	0.01	0.68	-3.28	-2.39	-1.94	-1.50	-0.55	4000.00
σ_β	2.41	0.01	0.57	1.60	2.01	2.32	2.69	3.76	4000.00
σ_γ	1.15	0.00	0.06	1.05	1.12	1.15	1.19	1.26	558.65
μ_{β_1}	-2.90	0.00	0.14	-3.18	-3.00	-2.90	-2.80	-2.63	1365.64
μ_{β_2}	-3.40	0.00	0.16	-3.71	-3.50	-3.40	-3.29	-3.08	1532.38
μ_{β_3}	-0.12	0.00	0.11	-0.35	-0.20	-0.13	-0.05	0.10	1232.40
μ_{β_4}	-2.33	0.00	0.13	-2.58	-2.41	-2.33	-2.25	-2.08	1319.83

Appendix E

Questionnaire used to measure professor performance

The original scale was 1 2 3 4 5. Categories 1,2 and 3 were merged. So, finally there was 3 categories for each questions.

(1) Does the professor attend classes regularly and punctually?

(2) Does the professor respect the agreed dates for academic activities, including evaluations and delivery of results?

(3) Does the professor prepare each of the sessions of the subject beforehand?

(4) Is the professor accessible and willing to provide academic help?

(5) Does the professor encourage group work, recognizing the successes and achievements in the learning activities?

(6) Does the professor demonstrate commitment and enthusiasm in their teaching activities?

(7) Do you consider that the topics seen in the subject met your expectations?

(8) Does the professor include learning experiences in places other than the classroom (workshops, laboratories, company, community, etc.)?

(9) Does the professor organize activities that allow me to exercise my oral and written expression?

(10) Does the professor develop the content of the class in an orderly and understandable manner?

(11) Does the professor promote self-study and research?

(12) Does the professor use technology (computer, videobeam, digital platforms, mail) as a means to facilitate student learning?

(13) Does the teacher promote the use of various digital tools to manage (collect, process, evaluate and use) information?

(14) Does the professor promote the safe, legal and ethical use of digital information?

(15) In general, the teacher's performance was:

Appendix F

Questionnaire used to measure subject relevance

The original scale was 1 2 3 4 5. Categories 1 and 2 were merged. So, finally there was 4 categories for each questions.

(1) Do you consider that the contents of the subject are clear and specific?

(2) Do you consider that the contents of the subject are related to that of others?

(3) Do you consider that the theoretical and practical contents of the subject are useful for the professional activity?

(4) How would you rate the subject in general terms?