EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 13/13

May 2013

Ranking Scientific Journals via Latent Class Models for Polytomous Item Response Data

by

Francesco Bartolucci

(University of Perugia)

Valentino Dardanoni

(University of Palermo)

Franco Peracchi

(University of Rome "Tor Vergata" and EIEF)

Ranking scientific journals via latent class models for polytomous item response data*

Francesco Bartolucci
University of Perugia

Valentino Dardanoni
University of Palermo

Franco Peracchi
Tor Vergata University and EIEF

May 16, 2013

Abstract

We propose a strategy for ranking scientific journals starting from a set of available quantitative indicators that represent imperfect measures of the unobservable "value" of the journals of interest. After discretizing the available indicators, we estimate a latent class model for polytomous item response data and use the estimated model to classify each journal. We apply the proposed approach to data from the Research Evaluation Exercise (VQR) carried out in Italy with reference to the period 2004–2010, focusing on the sub-area consisting of Statistics and Financial Mathematics. Using four quantitative indicators of the journals' scientific value (IF, IF5, AIS, h-index), some of which not available for all journals, we derive a complete ordering of the journals according to their latent value. We show that the proposed methodology is relatively simple to implement, even when the aim is to classify journals into finite ordered groups of a fixed size. Finally, we analyze the robustness of the obtained ranking with respect to different discretization rules.

KEYWORDS: Classification; Finite Mixture Models; Graded Response Model; Research Evaluation; VQR.

* We thank Sergio Benedetto, Tullio Jappelli, and Daniele Terlizzese for insightful discussions.

1 Introduction

There is a growing interest in issues surrounding the classification of scientific journals for evaluating research institutions or individual researchers. In fact, evaluation systems partially based on journal rankings have been recently introduced in various countries, for example in Australia by the Australian Research Council (ARC), in France by the "Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur" (AERES), and in Italy by the "Agenzia di Valutazione del Sistema Universitario e della Ricerca" (ANVUR).

There are by now many indicators which allow one to obtain a complete ordering of scientific journals, such as the Impact Factor (IF), the 5-year Impact Factor (IF5), the Article Influence Score (AIS), or the h-index, just to name a few, which are derived using commonly available databases such ISI-Thomson-Reuters, Scopus, or Google Scholar; see, among others, Garfield (2006), Bergstrom and West (2008), and Althouse et al. (2009). In a recent paper, Zimmermann (2012) describes 35 different indicators which can been used to rank journals. While different indicators may induce different rankings of a given set of journals, there is disagreement on whether there exists a single best general indicator and, if so, what this indicator is.

The aim of this paper is to propose a strategy for obtaining a unique ranking of scientific journals using a set of indicators of the value of a journal. There are several approaches for reducing a set of journal value indicators into a single ranking. One approach is Principal Component Analysis (PCA), which aims at extracting the latent value of each journal (e.g., Bollen et al., 2009). Another approach consists of taking some type of average of the rankings induced by the different indicators (e.g., the RePEc ranking of economic journals employs the harmonic mean of ranks after dropping the best and worst values). Relative to these approaches, the strategy that we propose has several advantages. First, it may be simply implemented on the basis of a meaningful statistical model. Second, it is able to produce a complete ordering of the journals. Third,

it provides a measure of the reliability of each indicator for classifying the journals in the chosen list. Finally, it can be applied with partially missing indicators' information.

Starting from the consideration that the scientific value of a journal is an unobservable (or latent) variable, we adopt a latent class version of the Graded Response Model (Samejima, 1969, 1996), which is commonly used in education for the analysis of polytomous item response data. After suitably discretizing the observed indicators, the model is estimated by maximum likelihood (ML) using readily available implementations of the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). Notice that our methodology is semiparametric in nature, since no assumption is made on the distribution of the unobserved latent value. Moreover, when a journal's unobserved value is intrinsically multidimensional, our procedure collapses the different dimensions into a unique underlying measure in a natural way, as it will be discussed in the sequel.

We apply our approach to data from the Research Evaluation Exercise ("Valutazione della Qualità della Ricerca" or VQR) carried out in Italy with reference to the period 2004–2010. This evaluation exercise involves all state universities, private universities granting publicly recognized academic degrees, and public research institutions. Researchers affiliated to these structures must submit for evaluation a number of research products (i.e., journal articles, books, book chapters, patents, etc.) published, or more generally produced, during the period 2004–2010. The typical number of products submitted by each researcher is three. The evaluation exercise is organized in 14 evaluation areas corresponding to broadly defined academic fields (e.g., Mathematics, Law, Economics and Statistics) and is carried out by a public agency (ANVUR) through Groups of Experts of Evaluation (GEV), one for each area. In most areas, journal articles are the main research products submitted to evaluation and, for each area, an important preliminary step is the ranking of the journals in which these articles have been published.

Using data on a number of quantitative indicators of the value of a journal – namely

the impact factor (IF), the 5-year impact factor (IF5), the article influence score (AIS), and the h-index – some of which are missing for certain journals, we derive a complete ordering of all the journals included in the list for the sub-area Statistics and Financial Mathematics. This sub-area belongs to the area Economics and Statistics and its products are evaluated by one of the GEVs, named hereafter GEV13. As we show through our application, the proposed methodology can handle missing data, is relatively simple to implement, and gives reasonable results. We discuss the robustness of the estimated ranking to different rule for discretizing the available indicators, and how to deal with the requirement that journals must be classified in ordered groups of *a priori* fixed size (e.g., this is indicated by the ANVUR guidelines).

The remainder of the paper is organized as follows. Section 2 describes the proposed ranking strategy. Section 3 presents the results obtained using the data from the Italian VQR 2004–2010 for the sub-area Statistics and Financial Mathematics. Finally, Section 4 provides some conclusions.

2 Proposed ranking strategy

Let n and r respectively denote the number of journals to be ranked and the number of indicators on which the ranking is to be based. In our case $r = 4$, since the available indicators are the IF, the IF5, and the AIS obtained from Thompson Reuters, plus the h-index obtained from Google Scholar. Also let x_{ij} be the value of indicator j for journal i, with $i = 1, \ldots, n$ and $j = 1, \ldots, r$. Note that the value of an indicator may be missing for some journals. In our data, this occurs for IF, IF5, and AIS, but never for the h-index.

Our strategy for ranking scientific journals is based on first discretizing the above indicators and then applying a statistical model for polytomous item response data. More precisely, let $q_{j1}, \ldots, q_{j,s-1}$ be a set of cutoffs or threshold values for the jth indicator x_{ij},

for example its quartiles or deciles, and define

$$y_{ij} = \sum_{m=0}^{s-1} m \cdot 1\{q_{jm} < x_{ij} \leq q_{j,m+1}\}, \qquad i = 1, \ldots, n, \tag{1}$$

where $q_{j0} = -\infty$, $q_{js} = \infty$, and $1\{A\}$ is the indicator function of the event A. Thus, y_{ij} is equal to 0 if $x_{ij} \leq q_{j1}$, is equal to 1 if $q_{j1} < x_{ij} \leq q_{j2}$, and so on until $y_{ij} = s - 1$ if $x_{ij} > q_{j,s-1}$. Clearly, if the value of x_{ij} is missing for some i and j, then the value of y_{ij} is also missing.

The main advantage of discretizing the available indicators, rather than working directly with the original values x_{ij}, is that we can use existing models with a straightforward interpretation and can rely on available software. Further, discretizing the observed indicators offers some robustness to measurement errors. However, since the way in which the available indicators are discretized is essentially arbitrary, it is important to asses the sensitivity of the results to the assumed discretization, as we will show in the application.

2.1 Statistical model

In this section we discuss our statistical model for the outcomes y_{ij} and show how to use it to predict the latent value of every journal in the given list. Our model is based on assumptions that typically characterize Item Response Theory (IRT) models (Hambleton and Swaminathan, 1985). We first consider the case where these outcomes are observable for all $i = 1, \ldots, n$ and $j = 1, \ldots, r$, so there is no missing data problem. We collect the r outcomes corresponding to the ith sample unit (i.e., journal) into the r-dimensional vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ir})$, where $y_{ij} = 0, \ldots, s - 1$ for all i and j.

The IRT model that we propose is based on the following assumptions:

1. For every sample unit $i = 1, \ldots, n$, the variables $y_{i1}, \ldots, y_{ir}$ are conditionally independent given a latent variable u_i.

2. The conditional distribution of every y_{ij} given u_i satisfies

$$\log \frac{p(y_{ij} \geq m|u_i)}{p(y_{ij} < m|u_i)} = \alpha_j(u_i - \beta_{jm}), \quad m = 1, \ldots, s-1, \tag{2}$$

   as in the Graded Response Model (Samejima, 1969).

3. The latent variables $u_1, \ldots, u_n$ are independent and have the same discrete distribution with k support points $\xi_1, \ldots, \xi_k$ and corresponding probabilities $\pi_1, \ldots, \pi_k$, with $\pi_h = p(u_i = \xi_h)$.

The first assumption, known as *local independence*, is typical of IRT models (Hambleton and Swaminathan, 1985). In the present context, it allows us to interpret the latent variable u_i as the intrinsic value of a journal, a latent construct which is the analog of the unobservable "ability" of an examinee in cases where the data are derived from the administration of test items. This assumption means that if we knew the value of u_i for the ith sample unit, then knowing the value of the jth indicator would not be useful to predict the value of any other indicator, since all the relevant information to capture the true value of a journal is already contained in u_i.

The second assumption formalizes our interpretation of the latent variable u_i. In particular, if the parameter α_j is positive, then the distribution of y_{ij} stochastically increases with u_i. In fact, parametrization (2) is based on the so-called *cumulative logits* (see Agresti, 2002, among others), which generalize the standard logits for binary outcomes to the case of ordinal outcomes. In practice, this means that the probability distribution of y_{ij} moves its mass towards higher classes as u_i increases. It is also worth noting that, in terms of the original outcomes x_{ij}, assumption (2) may equivalently be expressed as

$$\log \frac{p(x_{ij} \geq q_{jm}|u_i)}{p(x_{ij} < q_{jm}|u_i)} = \alpha_j(u_i - \beta_{jm}), \quad m = 1, \ldots, s-1.$$

In this regard, the parameter α_j, known in the IRT literature as the *discriminating index*, measures the sensitivity of the distribution of y_{ij} to changes in u_i, that in our context, is the

latent value of a journal. The interpretation of the parameters β_{jm} is context-specific. For example, in the educational context, they are interpreted as the levels of difficulty of the various items. As another interpretation of the model parameters, suppose that $x_{ij} = \gamma_j + \delta_j u_i + \varepsilon_{ij}$, where $\delta_j \neq 0$ and ε_{ij} is a zero-mean random variable distributed independently of u_i with a logistic distribution. Combining this model with the discretization rule (1) gives (2) with $\alpha_j = \delta_j$ and $\beta_{jm} = (q_{jm} - \gamma_j)/\delta_j$.

According to the third assumption, the distribution of each latent variable u_i is discrete. Since both the support points $\xi_1, \ldots, \xi_k$ and the corresponding probabilities $\pi_1, \ldots, \pi_k$ are parameters to be estimated, this assumption avoids specifying a parametric distribution for the latent variable. In this sense, our model is semiparametric in nature; see Lindsay et al. (1991) for a simpler semiparametric model for binary outcomes formulated along the same lines. As it will be clear in the following, if the aim is that of best approximating the distribution of u_i, then the number of support points k may be chosen on the basis of the observed data through a suitable selection criterion, such as the Bayesian Information Criterion (Schwarz, 1978). In other contexts, for instance when the size of each clusters is not constrained in advance, this number may be fixed a priori (see also the discussion at the end of Section 3).

Notice that the third assumption implicitly requires a journal's latent value to be unidimensional. While the unobserved value of a journal may have more than one dimension, as discussed for example by Bollen et al. (2009), undimensionality is a required assumption if we want to obtain a unique ranking of journals. Unidimensionality of the latent value may however be tested against multidimensionaly (see Bartolucci, 2007, among others).

Also notice that, by the third assumption, the latent variables $u_1, \ldots, u_n$ are also mutually independent, so the response vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ are independent across sample units. This independence assumption may restrictive in some cases, for example when the discretized outcomes y_{ij} are constructed using as cutoffs the sample quantiles, which

necessary depend on the overall distribution of the data. However, we expect that minor failures of this assumption should not significantly affect the results of the analyses based on the proposed approach, especially when the sample size n is large. On the other hand, relaxing this assumption would lead to a much more complex model.

Given our assumptions, the model parameters are the support points ξ_h and the corresponding probabilities π_h, $h = 1, \ldots, k$, the discriminant indices α_j, $j = 1, \ldots, r$, and the cutoffs β_{jm}, $j = 1, \ldots, r$, $m = 1, \ldots, s-1$. However, due to the identifying constraints $\alpha_1 = 0$ and $\beta_{11} = 0$ and the fact that $\sum_{h=1}^{k} \pi_h = 1$, the number of free parameters is only

$$\#\text{par}_k = k + (k-1) + (r-1) + [r(s-1) - 1] = 2k + rs - 3. \tag{3}$$

Notice that we can alternatively impose the identifying constraints, as we do in our application below, by standardizing the latent distribution to have zero mean and unit variance.

As already mentioned, our model is of the IRT type. In fact, it may be seen as a finite mixture version of the Graded Response Model, which is well known in the IRT literature (Samejima, 1969, 1996). The finite mixture nature of the model derives from considering the distribution of the latent variable as discrete.

Under the above three assumptions, the *manifest distribution* of $\boldsymbol{y}_i$ may be expressed as

$$p(\boldsymbol{y}_i) = \sum_{h=1}^{k} \pi_h \prod_{j=1}^{r} p(y_{ij} | u_i = \xi_h), \tag{4}$$

where $p(y_{ij}|u_i)$ is the conditional probability of the outcome y_{ij}, which satisfies (2). This manifest distribution is key for ML estimation of the model parameters.

It is also important to notice that the *posterior distribution* of u_i, namely the conditional distribution of u_i given $\boldsymbol{y}_i$, has the following probability mass function

$$p(u_i | \boldsymbol{y}_i) = \frac{\pi_h \prod_{j=1}^{r} p(y_{ij}|u_i)}{p(\boldsymbol{y}_i)}. \tag{5}$$

This probability is used to assign every sample unit to a given group (or latent class). In particular, once the model has been estimated, unit i is assigned to group h if

$$h = \underset{g=1,\ldots,k}{\operatorname{argmax}}\, p(u_i = \xi_g|\boldsymbol{y}_i). \tag{6}$$

Moreover, we can predict the value of u_i using the mean of the posterior distribution of u_i, or *posterior* mean, which is defined as follows

$$\hat{u}_i = \sum_{h=1}^{k} \xi_h p(u_i = \xi_h|\boldsymbol{y}_i). \tag{7}$$

When there are missing data, we compute the manifest distribution of the vector of observed outcomes as

$$p(\boldsymbol{y}_i) = \sum_{h=1}^{k} \pi_h \prod_{m_{ij}=0, j=1}^{r} p(y_{ij}|u_i = \xi_h),$$

where m_{ij} is an indicator variable equal to 1 if y_{ij} is missing and to 0 if it is observed. We rely on this expression for ML estimation. This amounts to assuming that the data are Missing-at-Random (MAR) in the sense of Little and Rubin (2002). In our context, MAR implies that the event that the value of an indicator – say IF5 – is missing may be predicted by the observable indicators, in our case the h-index. We consider this assumption realistic enough since missing values of certain indicators tend to be observed for journals with a lower reputation, as measured by the level of the h-index.

2.2 Likelihood inference

Given observations on a set of n journals, consisting of the discrete outcomes y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, r$, the sample log-likelihood is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\boldsymbol{y}_i),$$

where $\boldsymbol{\theta}$ is the vector containing all the model parameters and $p(\boldsymbol{y}_i)$ is the manifest probability of the response vector $\boldsymbol{y}_i$, computed according to (4) and depending on $\boldsymbol{\theta}$.

In order to maximize $\ell(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we use the version of the EM algorithm (Dempster et al., 1977) implemented as described in Bacci et al. (2012), to which we refer for details. This implementation is available in the `R` package `MultiLCIRT` (Bartolucci et al., 2013).

First of all, denoting by z_{hi} the (unobserved) indicator equal to 1 if $u_i = \xi_h$ and to 0 otherwise, the *complete* sample log-likelihood is equal to

$$\ell^*(\boldsymbol{\theta}) = \sum_{h=1}^{k} \sum_{i=1}^{n} z_{hi} \log \left[\pi_h \prod_{j=1}^{r} p(y_{ij}|u_i = \xi_h) \right]. \tag{8}$$

The EM algorithm alternates between the following two steps until convergence:

E-step: Compute the conditional expected value of $\ell^*(\boldsymbol{\theta})$ given the observed data and the current value of the parameters.

M-step: Maximize the above expected value with respect to $\boldsymbol{\theta}$ to get an updated estimate of the parameter vector.

The E-step consists of computing, for every h and i, the expected value of z_{hi} given $\boldsymbol{y}_i$ through the posterior probabilities in (5), and then substituting these expected values in (8). At the M-step, the resulting function is maximized with respect to $\boldsymbol{\theta}$. The existence of a closed-form solution for the estimates of the probabilities π_h makes the maximization problem easier, whereas updating the other parameters only requires simple iterative algorithms.

Finally, in applying the model we need a suitable criterion for choosing the number k of support points (or latent classes) of the distribution of u_i when this value is not *a priori* fixed. We refer to McLachlan and Peel (2000), Chapter 6, for an overview of the available criteria in the general context of finite mixture models, and to Dias (2006) for a more recent specific review of latent class models. To prevent selecting too many latent classes, we use the Bayesian Information Criterion (BIC) of Schwarz (1978), which is based on

minimization of the index

$$BIC_k = -2\ell(\hat{\boldsymbol{\theta}}_k) + \log(n)\,\#\text{par}_k, \tag{9}$$

where $\hat{\boldsymbol{\theta}}_k$ is the ML estimate of $\boldsymbol{\theta}$ under the model with k latent classes and $\#\text{par}_k$ is the corresponding number of parameters, which is defined in (3). The quality of the choice may be measured by the entropy index

$$E_k = -\sum_{i=1}^{n}\sum_{h=1}^{k} \hat{p}(u_i = \xi_h|\boldsymbol{y}_i)\log[\hat{p}(u_i = \xi_h|\boldsymbol{y}_i)],$$

which lies between 0 (perfect clustering) and $n\log(k)$.

3 Application

We apply our approach to the list of scientific journals for the sub-area Statistics and Financial Mathematics published by GEV13, available at the web page of ANVUR. The list was created starting from all journals in the ISI-JCR Social Science Edition of Thomson Reuters Web of Science that belong to the core subject categories for GEV 13. It also includes many journals in the ISI-JCR Science Edition that belong to subject categories which are considered relevant to the area. The initial list was expanded using the list of journals, provided by CINECA (a non-profit consortium formed by 54 Italian universities), in which at least one Italian researcher belonging to the area has published in 2004–2010.

To avoid different rankings across sub-areas, GEV13 assigned each journal in the list to one and only one of its four sub-areas (Business; Economics; Economic History; Statistics and Financial Mathematics). This led to excluding from the the list for the sub-area Statistics and Financial Mathematics of a very small number of journals assigned to other sub-areas based on their prevalent content. One example is *Econometrica*, assigned to the sub-area Economics.

For each of the $n = 445$ journals in the sub-area Statistics and Financial Mathematics, we have $r = 4$ indicators, namely the IF, the IF5, the AIS, and the h-index. Table 1 shows

descriptive statistics for these four indicators, the distribution of which is represented in Figure 1, Table 2 shows their correlation matrix, while Table 3 shows summaries of the distribution of the h-index for ISI and non-ISI journals, namely the minimum (Min), the lower quartile ($Q_{.25}$), the median ($Q_{.50}$), the mean, the upper quartile ($Q_{.75}$) and the maximum (Max). The large differences between the distribution of the h-index for ISI and non-ISI journals provide a strong justification for our MAR assumption.

3.1 Model fitting

As discussed at the beginning of Section 2, the first step of our journal ranking strategy consists of discretizing the observed indicators. We present two alternative discretizations: one uses as cutoffs the sample quartiles ($s = 4$), the other uses the sample deciles ($s = 10$). Given the discretized outcomes y_{ij}, we fit our model for increasing values of k. In particular, we increase the value of k until the BIC index (9) does not become smaller than that computed for the previous value of k. To avoid local maxima of the sample log-likelihood, following the current literature on latent class and finite mixture models we use two types of initialization (deterministic and random) of the EM algorithm. The results, for both $s = 4$ (quartiles) and $s = 10$ (deciles), are shown in Table 4.

Table 4 suggests two models for the data, one with $k = 4$ when $s = 4$ (Model 1) and one with $k = 7$ when $s = 10$ (Model 2). The entropy index for Model 1 is $E_4 = 135.71$, corresponding to 22.0% of the maximum value of 616.90, while for Model 2 is $E_7 = 213.48$, corresponding to 20.8% of the maximum value of 1024.65. The estimated distribution of the latent variable (support points and probabilities) under the two models is shown in Table 5. To facilitate the comparison between models and to provide a sensitivity analysis on the number of support points, the table also shows the distribution of the latent variable when $k = 4$ and $s = 10$ (Model 3). The support points are in increasing order, so they identify groups of journals of increasing value. Notice that we used the standardization of the latent trait as identifiability constraint.

When $k = 4$, the estimated distributions of the latent variable are rather similar using quartiles (Model 1) or deciles (Model 3). The two distributions are especially close in terms of estimated probabilities at every support point, with the probability of the four ordered groups of journals about equal to 50% for the first group, 20% for the second and third groups, and 10% for the last group (the best journals). The distribution when $k = 7$ (and $s = 10$) is not directly comparable with the distribution when $s = 4$. However, we can compare the overall rankings by examining the predicted values of u_i in the three cases, computed using (7). Figure 2 shows the scatterplot of the maximum posterior probabilities, while Figure 3 shows the scatterplot of the predicted latent values.

The Pearson (Spearman) correlation coefficients between the predicted latent values from Model 1 and Model 2 is equal to .974 (.944), between the predicted latent values from Model 1 and Model 3 is equal to .968 (.928), and between the predicted latent values from Model 1 and Model 3 is equal to .985 (.985). This suggests that, apart from rescaling, the predicted latent values from the three models are very close to each other and provide a very similar ranking of the journals.

3.2 Classification

We now compare the classification of the journals into four ordered categories which are obtained by the three different models, and contrast them with that provided by GEV13. GEV13 adopts a classification based on four groups of journals of size 216, 36, 81, and 112 respectively, so the relative weight of each group is close to that suggested by the VQR rules, namely 20%, 20%, 10% and 50%. However, if we use $k = 4$ and classify the journals on the basis of their maximum posterior probability – see expression (6) – we do not obtain groups with size equal to that used by GEV13, neither with $s = 4$ nor $s = 10$. Furthermore, fixing the probabilities π_h to values equal to the required proportions does not solve the problem, since the number of journals that are assigned to each class based on the maximum posterior probability can be very different from the target number.

To create classes of journals of the same size as the classification adopted by GEV13, we first order the journals according to the predicted value $\hat{u}_i$ of the latent value u_i. We then include the first 216 journal of the ordered list in the first class, the second 36 journal in the second class, and so on. Table 6 reports the corresponding cross-classifications, while Table 7 reports Cohen's κ index of agreement.

Overall, we observe a strong agreement between the classifications of the journals obtained under different values of s and k. In particular, the percentage of journals that change classification ranges from 7.9% (comparison between Model 2 and Model 3) to 11.5% (comparison between Model 1 and Model 3). The good agreement between the three models is confirmed by the values of Cohen's κ indix.

As for the comparison between these classifications and that set up by GEV13, the percentage of disagreement is somewhat higher and ranges from 18.4% (comparison with Model 2) to 22.0% (comparison with Model 1). We have to consider, however, that the classification produced by GEV13 does not use the IF as indicator, and uses the h-index alone as a predictor of IF5 and AIS when these indicators are missing.

3.3 Discriminant indices

The proposed approach also allows us to assess the quality of an indicator, as a measure of the latent value of a journal, using the estimates of the discriminant indices α_j in equation (2), reported in Table 8.

Notice that, as frequently happens with these models, some estimates take rather extreme values; the distribution of the estimator is actually known to be highly skewed. Table 9 presents the confidence intervals for the discriminant index from Model 1 ($s = 4, k = 4$), obtained through a parametric bootstrap, while Table 10 presents the confidence intervals for the ratios $\alpha_{j_2}/\alpha_{j_1}$, which may be used to compare the different bibliometric indicators. The null hypothesis $H_0 : \alpha_{j_2}/\alpha_{j_1} = 1$ that two indicators j_1 and j_2 have the same discriminant power may be tested by checking if 1 is contained in the confidence

interval.

Our results suggest that, at least in our sample, the IF5 has a significantly higher discriminant power than the other bibliometric indicators; while the AIS and the h-index do not have a significantly different discriminant power. This agrees with the results in Chang et al. (2010) who, using data from the ISI database of citations from all fields in Sciences and Social Sciences, conclude that the AIS does not add very much compared to more traditional indicators such as the IF5.

3.4 An alternative clustering model

The approach in the previous sections is based on discretizing the observed indicators x_{ij}, which are transformed into the categorical responses y_{ij}. In this section we compare this approach with an alternative approach based on directly modeling the distribution of the continuous indicators x_{ij} as a finite mixture.

The model that we consider can be described as follows:

- For every journal i, the latent variables $x_{i1}, \ldots, x_{ir}$ are conditionally independent given u_i.

- Given u_i, each variable x_{ij} satisfies the linear model

$$x_{ij} = \gamma_j + \delta_j u_i + \sigma_j \varepsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, r,$$

  where the ε_{ij} are independently identically distributed standard Gaussian errors.

- The latent variables $u_1, \ldots, u_n$ are independent and have the same discrete distribution with k support points $\xi_1, \ldots, \xi_k$ and corresponding probabilities $\pi_1, \ldots, \pi_k$.

This finite mixture of normal distributions may be fitted by a standard EM algorithm. We apply this model to both the original data (without any transformation of the indicators) and the transformed data (via a log transformation to account for skewness). BIC

leads to selecting $k = 6$ components with the original data and $k = 9$ components with the log-transformed data. Table 12 below shows the distribution of the latent variable for the two models with the chosen number of classes while Figure 4 shows the scatterplot of the predicted latent values. The Pearson and Spearman correlation coefficients between the predicted latent values from the two models are equal to .723 and .810 respectively.

We finally use the two estimated models to classify the journals into four classes with the same sizes as the GEV13 classes, namely 216, 36, 81, and 112 journals. In general, it appears that classifications are not very robust with respect to transformations of the response outcomes. In our sample, the normal mixture approach is rather sensitive to the parametric assumptions, a finding which is in agreement with the evidence in Shentu and Xie (2010) that discretizing continuous observations may increase the robustness of the model with respect to model misspecifications and contaminations.

4 Conclusions

We propose a method for ranking scientific journals based on a latent variable model for polytomous item responses. The latent variable, assumed to be discrete and interpreted as the unobservable "value" of a journal, is predicted on the basis of indicators, such as the IF, IF5, AIS and h-index, that are discretized to avoid strong parametric assumptions. We also show how to deal with missing values of some of these indicators.

The main advantage of our approach is that it relies on a well principled statistical model that has some nonparametric features. In particular, our approach does not require to specify a parametric model for the distribution of the latent variable representing a journal's value, which is instead treated as discrete with an arbitrary number of support points that identify groups of journals with similar characteristics. In practice, the number of groups is chosen on the basis of the observed data through a statistical criterion, such as BIC. Therefore, in a context of classification, we can decide a suitable number of groups

of homogenous journals in the light of the data. The method also provides an estimate of the size of each of these groups.

As an outcome of the proposed approach, the mean of the posterior distribution of the latent variable provides a prediction on a continuous scale of the latent value of each journal in the given list, so journals can be univocally ordered, the distance between any pair of journals can be compared, and journals can be classified into any arbitrary number of classes of a given size. Another relevant feature of the proposed approach is that it allows us to assess the discriminant power of each indicator, that is, the sensitivity and reliability of each indicator in the relationship with the latent value of a journal. For example, in the data we analyze we find that the IF5 appears to be the most reliable indicator of the value of a journal among the indicators that are used in this study.

It is important to recall that our approach is based on discretization of quantitative indicators, so the results of the analysis may depend on the choice of cutoffs adopted for this discretization. In an application, it is therefore important to asses the sensitivity of the results to the adopted discretization. As we shown in our application, which deals with the list of journals in the sub-area Statistics and Financial Mathematics of the Italian Research Evaluation Exercise, robustness can be checked by replicating the analysis with different discretizations and then comparing the results obtained. The results in our applications turn out to be fairly robust to different discretizations.

Our analyisis can be easily extended to handle the case where different indicators have different numbers of categories (so as to include, for instance, binary indicators), and to employ discriminant indices which are category-dependent. Further theoretical and empirical analysis could investigate the reliability of the estimates of the discriminant indices, given the high skewness in the bootstrap samples. A general comparison between our approach with the normal mixture model for the original outcomes is likely to be very useful. Finally, our method could be applied to other lists of journals in different fields.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York, 2nd edition.

Althouse, B. M., West, J. D., Bergstrom, T. C., and Bergstrom, C. T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, 60:27–34.

Bacci, S., Bartolucci, F., and Gnaldi, M. (2012). A class of multidimensional latent class irt models for ordinal polytomous item responses. Technical report, http://arxiv.org/abs/1201.4667.

Bartolucci, F. (2007). A class of multidimensional IRT models for testing unidimensionality and clustering items. *Psychometrika*, 72:141–157.

Bartolucci, F., Bacci, S., and Gnaldi, M. (2013). MultiLCIRT: Multidimensional latent class Item Response Theory models. R package version 2.3, URL http://CRAN.R-project.org/package=MultiLCIRT.

Bergstrom, C. and West, J. (2008). Assessing citations with the eigenfactor metrics. *Neurology*, 71:1850–1851.

Bollen, J., de Sompel, H. V., Hagberg, A., and Chute, R. (2009). A principal component analysis of 39 scientific impact measures. *PLoS ONE*, 4.

Chang, C.-L., McAleer, M., and Oxley, L. (2010). Journal impact factor versus eigenfactor and article influence. Technical Report KIER Working Papers 737, Kyoto University, Institute of Economic Research.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Dias, J. (2006). Model selection for the binary Latent Class model: A Monte Carlo simulation. In Batagelj, V., Bock, H.-H., Ferligoj, A., and Žiberna, A., editors, *Data Science and Classification*, pages 91–99. Springer, New York.

Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association*, 295:90–93.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Kluwer Nijhoff, Boston.

Lindsay, B., Clogg, C., and Greco, J. (1991). Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, 86:96–107.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, 17.

Samejima, F. (1996). Evaluation of mathematical models for ordered polychotomous responses. *Behaviormetrika*, 23:17–35.

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

Shentu, Y. and Xie, M. (2010). A note on dichotomization of continuous response variable in the presence of contamination and model misspecification. *Statistics in Medicine*, 29:2200–2214.

Zimmermann, C. (2012). Academic rankings with RePEc. Technical report, Federal Reserve Bank of St. Louis Working Paper 2012-023A, St. Louis, MO.

Table 1: Descriptive statistics for the observed indicators.

		IF	IF5	AIS	h-index
Missing values	(%)	43.8	52.6	52.6	0
Mean		1.056	1.472	.946	19.77
Variance		.418	.751	.480	267.59
Skewness index		1.325	1.526	1.938	1.575
Quartile	1st	.586	.840	.506	7.0
	2nd	.954	1.284	.721	14.0
	3rd	1.381	1.867	1.203	28.0
Decile	1st	.370	.590	.313	4.0
	2nd	.521	.766	.454	6.0
	3rd	.643	.967	.553	9.0
	4th	.754	1.108	.660	12.0
	5th	.954	1.284	.721	14.0
	6th	1.088	1.467	.871	19.0
	7th	1.257	1.741	1.026	24.0
	8th	1.561	2.132	1.362	32.0
	9th	1.906	2.513	1.892	42.6

Table 2: Correlation matrix of the observed indicators.

	IF	IF5	AIS	h-index
IF	1.000	.899	.693	.556
IF5	.899	1.000	.795	.579
AIS	.693	.795	1.000	.485
h-index	.556	.579	.485	1.000

Table 3: Summaries of the distribution of the h-index for ISI and non-ISI journals.

Journals	#	Min	$Q_{.25}$	$Q_{.50}$	Mean	$Q_{.75}$	Max
non-ISI	195	3	4.0	7.0	8.45	12.0	26
ISI (only IF)	39	1	9.0	14.0	14.82	17.5	46
ISI (IF, IF5 & AIS)	211	1	19.0	28.0	31.14	40.0	108

Table 4: Results from a preliminary model fit with $s = 4$ (quartiles) and $s = 10$ (deciles).

	$s = 4$			$s = 10$		
k	$\ell(\hat{\boldsymbol{\theta}}_k)$	$\#\text{par}_k$	BIC_k	$\ell(\hat{\boldsymbol{\theta}}_k)$	$\#\text{par}_k$	BIC_k
1	-1544.9	12	3163.0	-2564.3	36	5348.0
2	-1343.8	17	2791.2	-2347.7	41	4945.4
3	-1293.0	19	2702.0	-2271.2	43	4804.6
4	-1273.6	21	**2675.2**	-2233.2	45	4740.7
5	-1271.0	23	2682.3	-2216.8	47	4720.2
6				-2206.5	49	4711.9
7				-2197.2	51	**4705.5**
8				-2194.7	53	4712.7

Table 5: Estimated distribution of the latent variable for Model 1 ($s = 4, k = 4$), Model 2 ($s = 10, k = 4$) and Model 3 ($s = 10, k = 7$).

	Model 1		Model 2		Model 3	
h	$\hat{\xi}_h$	$\hat{\pi}_h$	$\hat{\xi}_h$	$\hat{\pi}_h$	$\hat{\xi}_h$	$\hat{\pi}_h$
1	-.840	.537	-.924	.478	-.997	.401
2	.288	.178	.169	.216	-.274	.156
3	1.092	.182	1.015	.194	.327	.123
4	1.941	.104	1.851	.113	.785	.123
5					1.229	.098
6					1.621	.058
7					2.208	.041

Table 6: Cross-classification of the journals in four groups (c1, c2, c3, c4) of size equal to that used by GEV13.

		Model 2				Model 3				GEV13			
		c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
Model 1	c1	208	1	7	0	203	10	3	0	189	17	10	0
	c2	8	22	6	0	13	16	7	0	27	5	4	0
	c3	0	13	65	3	0	10	67	4	0	14	54	13
	c4	0	0	3	109	0	0	4	108	0	0	13	99
Model 2	c1					211	5	0	0	197	13	6	0
	c2					5	21	10	0	19	12	5	0
	c3					0	10	70	1	0	11	56	14
	c4					0	0	1	111	0	0	14	98
Model 3	c1									200	12	4	0
	c2									14	11	10	1
	c3									2	13	52	14
	c4									0	0	15	97

Table 7: Cohen's κ index (standard errors in parentheses).

		Model 2	Model 3	GEV13
Model 1	unweighted κ	.8607	.8267	.6670
		(.0207)	(.0228)	(.0297)
	linear weights κ	.9211	.9113	.8225
		(.0126)	(.0123)	(.0173)
	quadratic weights κ	.9574	.9587	.9120
		(.0609)	(.0614)	(.0619)
Model 2	unweighted κ		.8913	.7214
			(.0185)	(.0278)
	linear weights κ		.9474	.8554
			(.0091)	(.0153)
	quadratic weights κ		.9780	.9312
			(.0609)	(.0618)
Model 3	unweighted κ			.7112
				(.0282)
	linear weights κ			.8488
				(.0157)
	quadratic weights κ			.9271
				(.0615)

Table 8: Estimated discriminant indices.

j	Model 1	Model 2	Model 3
1 IF	3.772	4.194	5.150
2 IF5	6.740	12.645	39.424
3 AIS	2.103	2.199	2.438
4 h-index	2.626	2.251	2.264

Table 9: Confidence intervals for the estimated discriminant indices for Model 1.

j	estimate	95%-interval	
1 IF	3.772	3.140	5.921
2 IF5	6.740	5.156	88.164
3 AIS	2.103	1.449	2.489
4 h-index	2.626	2.090	3.256

Table 10: Confidence intervals for the ratios $\alpha_{j_2}/\alpha_{j_1}$ for Model 1.

j_1	j_2	comparison	estimate	95%-interval	
1	2	IF5 vs. IF	1.787*	1.089	23.885
1	3	AIS vs. IF	.558*	.340	.660
1	4	h-index vs. IF	.696*	.414	.938
2	3	AIS vs. IF5	.312*	.018	.394
2	4	h-index vs. IF5	.390*	.031	.498
3	4	h-index vs. AIS	1.248	.928	1.803

Table 11: Preliminary fit for the original and the log-transformed data.

	original scale			log-transformed data		
k	$\ell(\hat{\boldsymbol{\theta}}_k)$	$\#\text{par}_k$	BIC_k	$\ell(\hat{\boldsymbol{\theta}}_k)$	$\#\text{par}_k$	BIC_k
1	-2612.2	8	5273.1	-1394.6	8	2838.0
2	-2396.9	13	4873.2	-1120.6	13	2320.4
3	-2309.5	15	4710.5	-1019.3	15	2130.1
4	-2268.0	17	4639.6	-942.7	17	1989.1
5	-2237.5	19	4590.8	-909.5	19	1934.9
6	-2217.1	21	**4562.3**	-880.8	21	1889.7
7	-2211.5	23	4563.3	-857.9	23	1856.1
8				-842.0	25	1836.4
9				-835.0	27	**1834.7**
10				-830.9	29	1838.7

Table 12: Estimated distribution of the latent variable for the original and the log-transformed data.

	original		log-transformed	
h	$\hat{\xi}_h$	$\hat{\pi}_h$	$\hat{\xi}_h$	$\hat{\pi}_h$
1	-.687	.539	-2.906	.008
2	.149	.259	-1.471	.217
3	.961	.109	-.504	.113
4	1.939	.069	-.063	.170
5	3.323	.017	.358	.156
6	5.321	.007	.730	.141
7			1.049	.095
8			1.396	.078
9			1.889	.023

Table 13: Agreement between journal classifications.

		log-transformed				GEV13			
		c1	c2	c3	c4	c1	c2	c3	c4
original	c1	187	14	15	0	173	22	20	1
	c2	28	0	8	0	28	1	6	1
	c3	1	22	57	1	15	13	40	13
	c4	0	0	1	111	0	0	15	97
log-transformed	c1					196	14	6	0
	c2					20	11	5	0
	c3					0	11	55	15
	c4					0	0	15	97

Table 14: Cohen's κ index (standard errors in parentheses).

		log-transformed	GEV13
original	unweighted κ	.6942	.5447
		(.0288)	(.0329)
	linear weights κ	.8258	.7173
		(.0684)	(.0671)
	quadratic weights κ	.9051	.8281
		(.0613)	(.0614)
log-transformed	unweighted κ		.7078
			(.0283)
	linear weights κ		.8488
			(.0690)
	quadratic weights κ		.9285
			(.0619)

Figure 1: Scatterplot of the observed indicators.

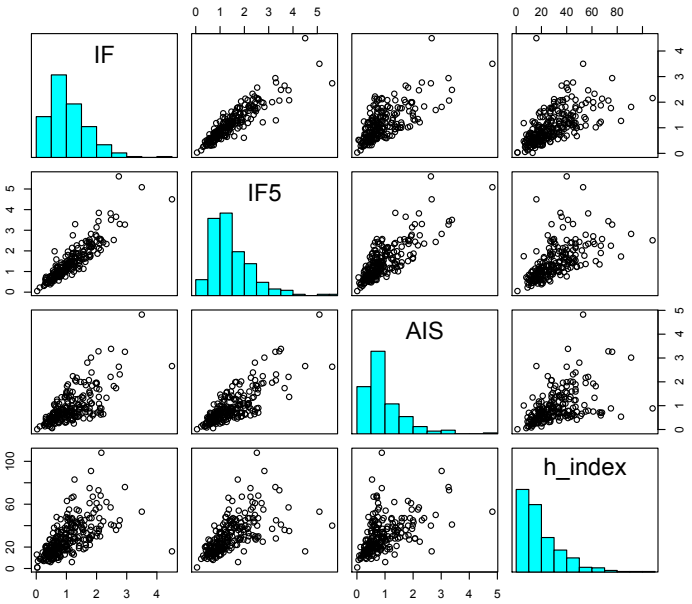

Figure 2: Scatterplot of the maximum posterior probabilities.

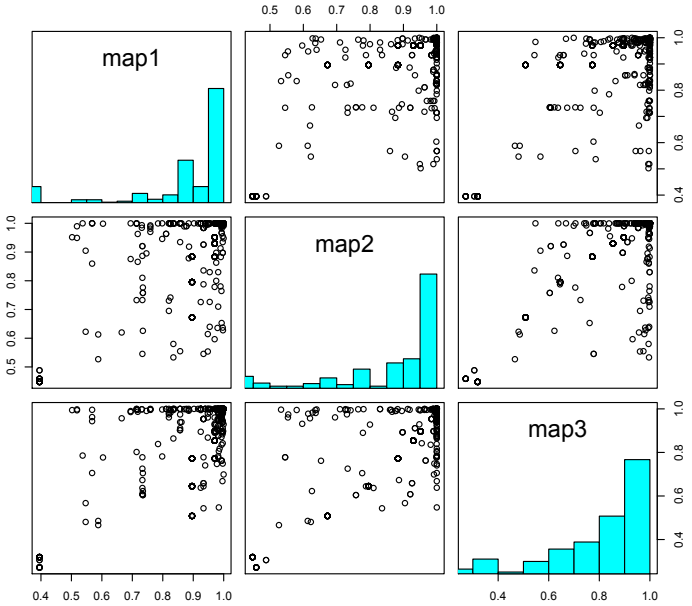

Figure 3: Scatterplot of the predicted latent values.

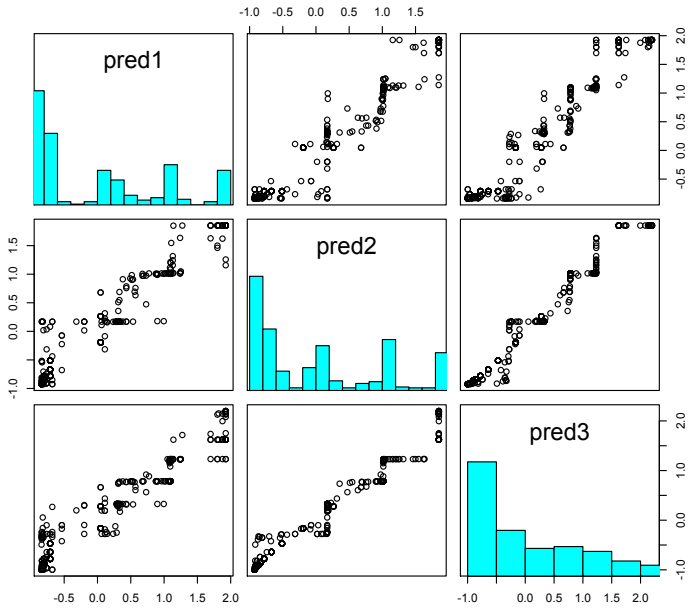

Figure 4: Scatterplot of the predicted latent values for the original (pred1) and the log-transformed (pred2) data.