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

Francesco Bartolucci
University of Perugia

Franco Peracchi
Tor Vergata University and EIEF

Valentino Dardanoni
University of Palermo

November 26, 2012

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 Applied Mathematics. Using four quantitative indicators of the journals' scientific value (IF, IF5, AIS, *h*-index), some of which are not observed for certain journals, we derive a complete ordering of the journals according to their latent scientific 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.

[*] Preliminary version. 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, such as Australia, by the Australian Research Council (ARC), France, by the "Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur" (AERES), and 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, e.g., Garfield (2006), Bergstrom and West (2008), and Althouse et al. (2009). In a recent paper, Zimmermann (2012) describes 35 different indicators which have been used to rank journals. While different indicators generally induce different rankings on a given set of journals, there is little agreement on whether there is a single best general indicator, and 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 quantitative indicators of the value of the journals in a chosen list. There are currently many practical approaches employed in this literature for reducing a set of of journal value indicators into a single ranking, such as using Principal Component Analysis (PCA), which extract the latent value of each journal (see e.g. (Bollen et al., 2009)), or 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 ranking). Our model: (*i*) may be simply implemented on the basis of a meaningful statistical model, (*ii*) is able to produce a complete ordering of the journals, and (*iii*) provides a measure of the reliability of each indicator for classifying the journals in the chosen list. 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 an 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 for 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; etc.) 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 ranking the journals in which these articles have been published.

Using data on a number of quantitative indicators of the scientific 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 Applied Mathematics, which is part of the area Economics and Statistics and whose 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 Applied Mathematics. Finally, Section 4 provides some conclusions.

2 Proposed ranking strategy

Let n denote the number of journals to rank and let r denote the number of indicators on which the ranking is to be based. In our case $r = 4$, since the indicators are the impact factor (IF), the 5-year impact factor (IF5) and the article influence score (AIS) obtained from Thompson Reuters, plus the h-index obtained from Google Scholar.[1] Also let x_{ij} be the value of indicator j for journal i, with $i = 1, \dots, n$ and $j = 1, \dots, r$. Note that the value of an indicator may be missing for some journals. In our data, this happens for IF, IF5 and AIS, but never for the h-index.

As outlined in the introduction, 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 $q_{j1}, \dots, q_{j,s-1}$ be a set of cutoffs or threshold values

[1] See http://www.isiknowledge.com/JCR for information on the properties of IF, IF5 and AIS, and http://scholar.google.com/intl/en/scholar/metrics.html for Google Scholar's h-index.

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 $x_{ij} > q_{j,s-1}$. The resulting outcomes y_{ij} are seen as discrete responses to items which are analyzed by the statistical model illustrated below. 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 we show how to use it to predict the latent scientific value of every journal in the given list. Our model is based on assumptions that typically characterize Item Response Theory (IRT) models. We first consider the case when 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 we propose to use 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 the parametrization

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

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 $\pi_1, \ldots, \pi_k$.

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 scientific value of a journal, a latent construct which is the analog of the unobservable 'ability' of an examinee in cases when 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, the 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.[2] 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.

[2] Recall that, given two discrete distributions for ordered variables with an equal support, one is stochastically larger than the other if and only if it has higher cumulative logits.

For example, in the educational context, they are interpreted as difficulty levels referred to the item categories.[3]

According to the third assumption, the distribution of each latent variable u_i is discrete, and, since both the support points $\xi_1, \ldots, \xi_k$ and the corresponding probabilities $\pi_1, \ldots, \pi_k$ are parameters to be estimated, it avoids the formulation of 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).

It is important to notice that the third assumption implicitly implies that a journal's latent value is unidimensional. While the unobserved value of a journal may have more than one dimension, as discussed for example by Bollen et al. (2009), since our purpose is to obtain a unique ranking of journals, unidimensionality is a required assumption. Unidimensionality of the latent value may however be tested against multidimensionaly; see e.g. Bartolucci (2007).

Finally notice that, according to 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

[3] To clarify the 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 $\beta_{jm} = (q_{jm} - \gamma_j)/\delta_j$.

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, which is beyond the scope of the present paper.

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 γ_j, $j = 1, \ldots, r$, and the cutoffs β_{jm}, $j = 1, \ldots, r$, $m = 1, \ldots, s-1$. However, since $\sum_{h=1}^{k} \pi_h = 1$ and due to the identifiability constraints $\gamma_1 = 0$ and $\beta_{11} = 0$, the number of free parameters is only

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

As already mentioned, the model we formulate 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; see Bacci et al. (2012) for further details.

The above assumptions imply that 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 $\pi_h = p(u_i = \xi_h)$ is the probability of the hth support point of the distribution of the latent variable and $p(y_{ij}|u_i)$ is the conditional probability of the outcome y_{ij}, which satisfies (2). This manifest distribution is key for ML estimation of the model parameters, as will be clarified below.

It is also important to recall 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

$$p(u_i = \xi_h|\boldsymbol{y}_i) = \max_{g=1,\ldots,k} p(u_i = \xi_g|\boldsymbol{y}_i). \tag{6}$$

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 outcome as

$$p(\boldsymbol{y}_i) = \sum_{h=1}^{k} \pi_h \prod_{j=1}^{r} [p(y_{ij}|u_i = \xi_h)d_{ij} + (1 - d_{ij})],$$

where d_{ij} is an indicator variable equal to 1 if y_{ij} is observed and to 0 otherwise, and $p(y_{ij}|u_i = \xi_h)$ is set equal to an arbitrary value when y_{ij} is missing. 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 rather realistic since missing values of certain indicators tend to be observed for journals with a lower reputation and a lower level of the h-index. On the other hand, in our application the h-index is always observable, so it is sensible to take it into consideration as a predictor when other indicators are missing.

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

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

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 a version of the EM algorithm (Dempster et al., 1977), implemented as described in Bacci et al. (2012) to which we refer for details. In our application, we rely on the implementation of this algorithm available in the `R` package `MultiLCIRT` (Bartolucci et al., 2012).

First of all, denoting by z_{hi} the (unobserved) indicator variable 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 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 $\boldsymbol{\theta}$. Regarding the probabilities π_h we have an explicit solution that may be used for this maximization, whereas updating the other parameters requires simple iterative algorithms.

Finally, in applying the model we need a suitable criterion to choose the number of support points (or latent classes) of the distribution of u_i, denoted by k, when this value is *a priori* fixed. [4] To prevent selecting too many latent classes, we use the Bayesian

[4] For an overview of the available criteria in the general context of finite mixture models, see McLachlan and Peel (2000), Chapter 6. For a more recent review specifically referred to latent class models, see Dias (2006).

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 maximum likelihood 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).

3 Application

To illustrate our approach, we consider the list of scientific journals for the sub-area Statistics and Applied Mathematics published by GEV13, available at the ANVUR web page.[5] The list was created starting from all journals in the ISI-JCR Social Science Edition of Thomson Reuters Web of Science (from now on WoS) that belong to the core subject categories for GEV13. 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 U-Gov 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, each journal in the list has been assigned to one and only one of the four sub-areas covered by GEV13 (Business; Economics; Economic History; Statistics and Applied Mathematics).

The list for the sub-area Statistics and Applied Mathematics contains $n = 445$ journals excluding a very small set of journals (for example, *Econometrica*) attributed to other sub-areas (Business, Economics, or Economic History). For each journal, the list includes $r = 4$ indicators, namely the IF, the IF5, the AIS and the h-index. Table 1 shows the percentage of missing values for each of the above indicators, and the quartiles and deciles of their observed distributions.

As discussed at the beginning of Section 2, the first step of our journal ranking strategy

[5] `http://www.anvur.org/?q=it/content/lista-riviste`.

		IF	IF5	AIS	h-index
Missing values	(%)	43.8	52.6	52.6	
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 1: *Descriptive statistics for the observed dataset.*

consists of discretizing the observed values of the 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}, obtained through (1), we fit our model for increasing values of k. In particular, we increase the value of k until the BIC, defined in (9), does not become smaller than that computed for the previous value of k. The results, for both $s = 4$ (quartiles) and $s = 10$ (deciles), are shown in Table 2. To prevent 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.

	$s = 4$ (quartiles)			$s = 10$ (deciles)		
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 2: *Results from a preliminary fit of the model with both $s = 4$ and $s = 10$.*

Table 2 suggests that a suitable number of support points is $k = 4$ when $s = 4$ and

$k = 7$ when $s = 10$. The corresponding estimated distribution of the latent variables (support points and probabilities) is shown in Table 3, where we also add the results for $k = 4$ when $s = 10$ in order to facilitate the comparison with the other cases and provide a sensitivity analysis on the number of support points. The table reports the support points in increasing order, so they identify groups of journals with increasing value.

	$s = 4$		$s = 10$ ($k = 4$)		$s = 10$ ($k = 7$)	
k	$\hat{\xi}_h$	$\hat{\pi}_k$	$\hat{\xi}_h$	$\hat{\pi}_k$	$\hat{\xi}_h$	$\hat{\pi}_k$
1	-1.391	.537	.344	.478	-0.139	.401
2	2.863	.178	4.929	.216	3.585	.156
3	5.898	.182	8.474	.194	6.680	.123
4	9.099	.104	11.983	.113	9.042	.123
5					11.328	.098
6					13.347	.058
7					16.367	.041

Table 3: *Estimated distribution of the latent variable when $s = 4$ (with $k = 4$) and $s = 10$ (with $k = 4$ and $k = 7$).*

When $k = 4$, the estimated distributions of the latent variable are rather similar using quartiles or deciles. The two distributions are especially close in terms of estimated probabilities at every support point. Four ordered groups of journals are found and their size is about equal to 50% for the first group, 20% for the second and third group, and 10% for the last group (the best journals). It is interesting to note that these percentages are close to the ones suggested by the Italian VQR except that, in the VQR, the class of best journals is expected to have size 20% and the class of medium level journals (the third class) is expected to have size 10%.

The distribution when $s = 10$ and $k = 7$ is not directly comparable with the previous ones. However, we can compare the overall rankings induced by the two different discretizations in terms of (Pearson or Spearman) correlation coefficients between the corresponding predicted values of u_i, which are computed by (7) for the three cases. The results are shown in Table 4.

The results in Table 4 suggest that, apart from rescaling, the predicted values of the

	Pearson			Spearman		
	pred1	pred2	pred3	pred1	pred2	pred3
pred1	1.000	.974	.968	1.000	.944	.928
pred2	.974	1.000	.985	.944	1.000	.985
pred3	.968	.985	1.000	.928	.985	1.000

Table 4: *Pearson and Spearman correlation coefficients between the predicted values of the latent variable for* $s = 4$ *and* $k = 4$ *(pred1),* $s = 10$ *and* $k = 4$ *(pred2), and* $s = 10$ *and* $k = 7$ *(pred3).*

latent variables are very similar and provide a very similar ranking of the journals. In this regard, it is also useful to compare the classification of the journals under the different models and contrast it with that provided by GEV13. In particular, 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.

It is worth noting that if we use $k = 4$ and we 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 $s = 10$. Also note that fixing the probabilities π_h to values equal to the required proportions does not solve the problem because 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 proceed as follows: (i) we order the journals according to the predicted value $\hat{u}_i$ of the latent value u_i; (ii) we include the first 216 journal of the ordered list in the first class, the second 36 journal in the second class, and so on. In order to evaluate the agreement between the different rankings, we report in Table 5 the corresponding cross-classifications.

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 $k = 4$ and $k = 7$, with $s = 10$ in both cases) to 11.5% (comparison between $s = 4$ with $k = 4$ and $s = 10$ with $k = 7$). As

		$s = 10$ ($k = 4$)				$s = 10$ ($k = 7$)				GEV13			
		c1	c2	c3	c4	c1	c2	c3	c4	c1	c2	c3	c4
$s = 4$	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
$s = 10$	c1					211	5	0	0	197	13	6	0
($k = 4$)	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
$s = 10$	c1									200	12	4	0
($k = 7$)	c2									14	11	10	1
	c3									2	13	52	14
	c4									0	0	15	97

Table 5: *Cross-classification of the statistical journals in four groups (c1, c2, c3, c4) having the same size used by GEV13.*

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 $s = 10$ and $k = 4$) to 22.0% (comparison with $s = 4$ and $k = 4$). We have to consider, however, that the classification produced by GEV13 does not use IF among the indicators, and uses the h-index alone as a predictor of IF5 and AIS when these indicators are missing.

Finally, it is worth noting that the proposed approach also allows us to assess the quality of an indicator as a measure of the latent scientific value of a journal. This assessment is based on the estimates of the discriminant indices (see equation (2)) that, for the present dataset, are reported in Table 6.

j	$s = 4$	$s = 10$ ($k = 4$)	$s = 10$ ($k = 7$)
1 (IF)	1.000	1.000	1.000
2 (IF5)	1.784	3.015	7.655
3 (AIS)	.558	.524	.473
4 (h-index)	.696	.537	.440

Table 6: *Estimates of the discriminant indices* (γ_j) *with* $s = 4$ ($k = 4$) *and* $s = 10$ ($k = 4$ *and* $k = 7$).

Our results show that, at least when assessing the quality of a scientific journal in this field, IF5 seems to be more reliable than IF. On the other hand, the estimated value of the discriminating index for AIS, which is lower than that for IF, is perhaps surprising since AIS is an index computed by a much more elaborated method. In this regard, Chang et al. (2010), using data from the ISI database of citations from all fields in Sciences and Social Sciences, concludes that AIS does not add very much compared to more traditional indicators, such as IF5. It is also worth noting that this estimate is close to that for the h-index.

4 Conclusions

We propose a method to rank scientific journals based on a latent variable model for the analysis of 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 IF, IF5, AIS and h-index, that are suitably discretized (e.g. on the basis of the quantiles of their observed distributions). 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 model that has some non-parametric features. In particular, our approach does not require to specify a parametric model for the distribution of the latent variable representing the value of a journal, that is instead treated as discrete with an arbitrary number of support points which 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 have a well principled method to decide what is the suitable number of groups of homogenous journal to be used 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 and the distance between any pair of journals can be evaluated. This also allows us to classify journals in any arbitrary number of classes of a given size. This is illustrated in some details in our application, which deals with the list of journals in the sub-area Statistics and Applied Mathematics provided by the Italian Group of Experts of Evaluation for Economics and Statistics (GEV13).

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 IF5 appears to be the most reliable indicator of the value of a journal among the indicators that are used in the study.

Finally, it is important to recall that our approach is based on discretization of quantitative indicators, so the results of an 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 shown in our application, this may be done by replicating the analysis with different discretizations and then comparing the results obtained. A sensitivity analysis may also be carried out with respect to the number of support points or latent classes, although using an information criterion may represent a better alternative.

References

Agresti, A. (2002). *Categorical Data Analysis (2nd Edition)*. John Wiley & Sons, Hoboken.

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. (2012). MultiLCIRT: Multidimensional latent class Item Response Theory models. R package version 1.0, 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, 2nd edition.

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

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.

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