ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2016)
178, *Part* 4, *pp.* 1025–1049

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

Francesco Bartolucci,

University of Perugia, Italy

Valentino Dardanoni

University of Palermo, Italy

and Franco Peracchi

University of Rome Tor Vergata and Einaudi Institute for Economics and Finance, Rome, Italy

[Received November 2013. Final revision November 2014]

Summary. We propose a model-based strategy for ranking scientific journals starting from a set of observed bibliometric indicators that represent imperfect measures of the unobserved 'value' of a journal. After discretizing the available indicators, we estimate an extended latent class model for polytomous item response data and use the estimated model to cluster journals. We illustrate our approach by using the data from the Italian research evaluation exercise that was carried out for the period 2004–2010, focusing on the set of journals that are considered relevant for the subarea statistics and financial mathematics. Using four bibliometric indicators (IF, IF5, AIS and the *h*-index), some of which are not available for all journals, and the information contained in a set of covariates, we derive a complete ordering of these journals. We show that the methodology proposed is relatively simple to implement, even when the aim is to cluster journals into a small number of ordered groups of a fixed size. We also analyse the robustness of the obtained ranking with respect to different discretization rules.

Keywords: Clustering; Finite mixture models; Graded response model; Item response theory models; Research evaluation; *Valutazione della Qualità delle Ricerca*

1. Introduction

There is growing interest in issues surrounding the classification of scientific journals as an intermediate step towards evaluating research institutions or individual researchers. In fact, evaluation systems partially based on journal rankings have recently been introduced in various countries, e.g. in Australia by the Australian Research Council, in France by the Agence d'Évaluation de la Recherche et de l'Enseignement Supérieur and in Italy by the Agenzia di Valutazione del Sistema Universitario e della Ricerca. Since many of these evaluation systems aim at clustering journals into merit classes, the problem arises of how many classes should be considered and how they should be constructed given the available information.

A large number of bibliometric indicators is now available, typically derived from either citation or usage log data, and any of them can in principle be employed to rank journals. Recently

Address for correspondence: Francesco Bartolucci, Department of Economics, Finance and Statistics, University of Perugia, Via A. Pascoli 20, Perugia 06123, Italy.
E-mail: bart@stat.unipg.it

 0964–1998/15/1781025

Zimmermann (2012) considered seven indicators, five of which—four different versions of the impact factor (see, for example, Garfield (2006)) plus the h-index (Hirsch, 2005)—are based on alternative transformations of the number of citations that a journal receives, and the other two are based on the number of abstract views and downloads. Chang *et al.* (2012) considered a larger set of 15 indicators for 299 leading international journals in economics. Bollen *et al.* (2009) considered an even larger set of 39 indicators that include traditional measures based on citation counts and other measures based on social network analysis and usage log data. Many of these bibliometric indicators are now routinely computed in commonly available citation databases such as the ISI–Thomson Reuters *Web of Science*, SciVerse Scopus, Google Scholar or Microsoft Academic Search, which helps to explain their widespread use.

Although rankings of journals generally differ depending on what indicator or database is used, there is disagreement on whether there is a single best general indicator or, more generally, on how to combine the information that is contained in the available indicators to rank journals. One approach is principal component analysis, which aims at extracting from the data the orthogonal components that account for most of the observed correlation between the indicators considered (see, for example, Bollen *et al.* (2009)). The problem with principal component analysis is that interpretation is not simple, especially when more than one component is needed to account for a significant fraction of total variance, so the association of the components with elusive concepts such as 'value', 'popularity' or 'prestige' of a journal remains rather speculative. An alternative approach is to take some kind of average of the rankings induced by the different indicators. For example, the 'Research papers in economics' ranking of economic journals (which is available from `https://ideas.repec.org/top/top.journals.all.html`) employs the harmonic mean of ranks after dropping the best and worst values. An important drawback of both approaches is that they are based neither on an explicit model for measuring intellectual influence (Palacios-Huerta and Volij, 2004) nor on a well-defined statistical model, which makes it difficult to interpret the results that are obtained and to carry out inference.

In this paper we propose a model-based strategy for ranking scientific journals starting from a set of observed bibliometric indicators that represent imperfect measures of the unobserved scholarly influence of a journal, treated initially as a unidimensional latent trait. After discretizing these indicators to avoid strong parametric assumptions, we estimate several latent class models for polytomous item response data and use them to cluster journals. Our models are latent class versions of the graded response model (Samejima, 1969, 1996), which is commonly used in educational research and is one of the most popular item response theory (IRT) models (Hambleton and Swaminathan, 1985). Relative to other approaches, our strategy offers several advantages. First, it is based on a well-defined statistical model and is easy to implement. Second, it produces a complete ordering of the journals. Third, it is semiparametric in nature, since it requires no parametric assumption on the distribution of the latent traits. Fourth, it provides a measure of the reliability of each indicator for classifying or clustering journals. Finally, it can be applied even when the bibliometric indicators contain missing values.

We illustrate our approach by using data from the recent Italian research evaluation exercise ('*Valutazione della Qualità della Ricerca*' (VQR)). The VQR involved all state universities, all private universities granting publicly recognized academic degrees and several public research centres. Researchers affiliated to these institutions were asked to submit for evaluation a number of research products published during the period 2004–2010 and participating institutions were then ranked on the basis of the average score received by the products submitted by their researchers.

The VQR was co-ordinated by a public agency (the Agenzia di Valutazione del Sistema Universitario e della Ricerca) through groups of experts ('*Gruppi di Esperti della Valutazione*'

(GEV)), one for each of 14 broadly defined scientific areas. In most areas, journal papers were the main type of research products submitted for evaluation. For these areas, an important input to the evaluation process was the ranking of the journals where these papers were published. In our application, we focus on the set of journals that were considered relevant for the subarea statistics and financial mathematics, as defined in the VQR. We discuss the robustness of the estimated journal ranking to different rules for discretizing the available indicators, and how to handle the VQR requirement that journals must be classified in ordered groups of *a priori* fixed size.

To be as close as possible to the actual problems that are faced in the VQR, we base our ranking of journals on exactly the same bibliometric indicators as used by GEV 13, namely the 2-year and 5-year impact factors IF and IF5, the article influence score AIS and the h-index. IF and IF5 measure the average number of citations to papers published in a given journal during the 2 and 5 previous years respectively. AIS uses the same information as IF5, but it also considers which journals have contributed to these citations (highly cited journals influence the indicator more than lesser cited journals) and removes journal self-citations. A journal has an h-index of h if h of the N papers that it published during a given period receive at least h citations each, and the other $N-h$ receive no more than h citations each.

The remainder of the paper is organized as follows. Section 2 describes the ranking strategy proposed, focusing in particular on the statistical models on which the strategy is based. Section 3 deals with estimation of these models and selection of the proper model specification. Section 4 illustrates our methodology by presenting the results that were obtained by using the data from the Italian VQR. Finally, Section 5 provides some conclusions.

2. Ranking strategy proposed

Let n denote the number of journals to be ranked and let r denote the number of available indicators on which the ranking is to be based. In our application (Section 4), the available indicators are IF, IF5 and AIS obtained from the *Web of Science* plus the h-index obtained from Google Scholar, so $r=4$. 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 application this occurs for IF, IF5 and AIS, but never for the h-index. Our strategy for ranking scientific journals is based on a preliminary discretization of the available indicators in s categories. A statistical model for polytomous item response data is then applied to the discretized indicators.

Denoting by $q_{j1},\ldots,q_{j,s-1}$ the cut-offs or threshold values for the jth indicator (e.g. its quartiles), the discretized value of indicator j for journal i is defined as

$$y_{ij}=\sum_{v=0}^{s-1} v\mathbf{1}\{q_{jv}<x_{ij}\leqslant q_{j,v+1}\}, \qquad i=1,\ldots,n, \tag{1}$$

where $q_{j0}=-\infty$, $q_{js}=\infty$ and $\mathbf{1}\{A\}$ is the indicator function of the event A. Thus, y_{ij} is equal to 0 if $x_{ij}\leqslant q_{j1}$, is equal to 1 if $q_{j1}<x_{ij}\leqslant q_{j2}$, 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. We collect the r indicators corresponding to the ith journal in the r-dimensional vector $\mathbf{y}_i=(y_{i1},\ldots,y_{ir})$.

Discretizing the available indicators, rather than working directly with their original values, implies some loss of information but allows us to avoid strong parametric assumptions. It also 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 discretization assumed.

In Section 2.1 we first describe a baseline model, relying on typical IRT assumptions, to analyse the data that were described above. These assumptions may be relaxed, giving rise to the extended models that are described in Section 2.2. Then, in Section 2.3, we show how to use these models to cluster journals into ordered groups and to predict their latent scientific impact.

2.1. Baseline item response theory model

Our baseline IRT model assumes the existence of n latent variables $u_1, \ldots, u_n$, which we interpret as the scholarly impact of each journal. In interpreting these latent variables we may take into account that IF, IF5 and AIS are measures of the immediate impact of a journal (respectively 2 and 5 years after publication for IF and IF5, and 5 years after for AIS), whereas the h-index may be seen as a measure of lifetime impact. Consequently, in this baseline IRT model, a journal has a high impact (i.e. a high value of u_i) if it has both a high immediate impact and a high lifetime impact. Distinguishing between these two dimensions may be important but would require two latent variables and, therefore, a bidimensional model of the type that is described in detail in Section 2.2. Note, however, that, whereas the unobserved impact of a journal may be multidimensional, as argued for example by Bollen *et al.* (2009), unidimensionality is essential if we want to obtain a unique ranking of journals.

Our baseline IRT model relies on the following three assumptions.

Assumption 1. For every sample unit (i.e. journal) $i = 1, \ldots, n$, the discretized values $y_{i1}, \ldots, y_{ir}$ are conditionally independent given the latent variable u_i.

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

$$\log\left\{\frac{p(y_{ij} \geqslant v|u_i)}{p(y_{ij} < v|u_i)}\right\} = \alpha_j(u_i - \beta_{jv}), \qquad v = 1, \ldots, s-1, \tag{2}$$

as in the graded response model (Samejima, 1969).

Assumption 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 $\pi_h = p(u_i = \xi_h)$. We then denote by $\lambda_{j|h}(v) = p(y_{ij} = v|u_i = \xi_h)$ the probability of the vth category for the jth discretized indicator y_{ij} conditionally on the latent variable u_i taking value ξ_h. Obviously, these conditional probabilities satisfy condition (2).

Assumption 1, which is known as *local independence*, is typical of IRT models (Hambleton and Swaminathan, 1985). It also characterizes latent class models (Lazarsfeld and Henry, 1968; Goodman, 1974) in which the conditional response probabilities $\lambda_{j|h}(v)$ are not constrained to take a specific parametric form but are free parameters to be estimated. In practice, this assumption means that, if we knew the value of u_i, then knowing the value of one indicator would not be useful to predict the value of any other indicator, as all the relevant information to capture the latent impact of a journal is already contained in u_i.

Assumption 2 formalizes our interpretation of the latent variable u_i. In particular, if the parameter α_j, which is known in the IRT literature as the *discriminating index*, is positive, then the distribution of y_{ij} stochastically increases with u_i. In fact, parameterization (2) is based on a version of cumulative (or global) logits (see Agresti (2002) among others), which generalize the standard logits for binary outcomes to the case of ordinal outcomes. Therefore, assumption 2 means that the probability distribution of y_{ij} moves its mass towards higher categories as u_i increases. It is worth noting that, in terms of the original indicators x_{ij}, it may equivalently be expressed as

$$\log\left\{\frac{p(x_{ij} \geqslant q_{jv}|u_i)}{p(x_{ij} < q_{jv}|u_i)}\right\} = \alpha_j(u_i - \beta_{jv}), \qquad v = 1, \ldots, s-1.$$

This shows that the parameter α_j measures the sensitivity of the distribution of x_{ij} to changes in u_i that, in our context, represents the latent impact of a journal. The interpretation of the parameters β_{jv}, which are known in the IRT literature as the *difficulty parameters*, is context specific. In fact, in the educational context they are interpreted as the levels of difficulty of the various test item categories. In the present context, β_{jv} is interpreted as a measure of severity in reaching category v for indicator j, in the sense that, with $\alpha_j > 0$, the probability that journal i falls in category v or in a higher category is larger than the probability that it falls in one of the previous categories if and only if $u_i > \beta_{jv}$. Consequently, these difficulty parameters are in increasing order, i.e. $\beta_{j1} < \ldots < \beta_{j,s-1}$, for all j. For another interpretation of the model parameters, suppose that the original indicators satisfy the linear model $x_{ij} = \gamma_j + \delta_j u_i + \varepsilon_{ij}$, 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 linear model with the discretization rule (1) gives model (2) with $\alpha_j = \delta_j$ and $\beta_{jv} = (q_{jv} - \gamma_j)/\delta_j$.

According to assumption 3, the latent variables $u_1, \ldots, u_n$ follow the the same discrete distribution. 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 the need to specify a parametric distribution for the latent variables. Thus, our model is semiparametric in nature; see Lindsay *et al.* (1991) for a simpler semiparametric model for binary outcomes formulated along the same lines.

Assuming a discrete distribution for the latent variables is quite natural if the aim of the model is to cluster journals, giving rise to a model-based clustering (Fraley and Raftery, 2002). In this case, k may be fixed *a priori*, provided that the size of each cluster is not constrained in advance. When not fixed *a priori*, k may be chosen through some statistical criterion, such as the Bayesian information criterion BIC that was introduced by Schwarz (1978).

Assuming a discrete distribution for the latent variables makes sense even if we believe that their distribution is continuous. In fact, the discrete distribution assumed is better seen as a convenient approximation to the unknown continuous distribution, the quality of the approximation increasing with k. In this regard, it has been established that even moderate values of k may provide an adequate approximation and even a better fit with respect to alternative models based on a parametric continuous distribution (Heckman and Singer, 1985; Lindsay *et al.*, 1991).

By assumption 3 the latent variables $u_1, \ldots, u_n$ are mutually independent, so the response vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are also independent across sample units. This independence assumption may be restrictive in some cases, e.g. when the discretized indicators y_{ij} are constructed by using as cut-offs a set of sample quantiles, which necessarily depend on the joint distribution of the data. However, such minor failures are unlikely to matter much, especially when the sample size n is large. In contrast, relaxing the independence assumption would lead to a much more complex model.

Given the above three 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 difficulty parameters β_{jv}, $j = 1, \ldots, r$, $v = 1, \ldots, s-1$. However, because of the identifiability constraints $\alpha_1 = 1$ and $\beta_{11} = 0$ and the fact that $\Sigma_{h=1}^{k} \pi_h = 1$, the number of free parameters is

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

To make the model identifiable we can alternatively impose, as we do in our application, that the latent distribution has zero mean and unit variance. This identifiability constraint is equivalent to that directly expressed on the parameters ($\alpha_1 = 1, \beta_{11} = 0$), in the sense that the same maximum of the likelihood is reached under both constraints. However, fixing the mean and the variance of the latent distribution makes the interpretation of the results easier and helps in comparing results that are obtained under different model specifications, as will be clear in our application (see in particular Section 4.4).

2.2. Extended item response theory models

In our application to the Italian VQR, the assumptions of the baseline IRT model may be too restrictive. In this section we show how to relax these assumptions, while retaining the main features of the proposed IRT model in terms of interpretability.

Local independence (assumption 1) is certainly restrictive, as some of the observed indicators (in particular IF and IF5) are constructed starting from the same information and using a similar method. Relaxing this assumption requires allowing for dependence between the indicators even after conditioning on the latent journal impact. This in turn raises the issue of how to model the conditional association between the indicators. For this aim, we adopt a Plackett formulation (Plackett, 1965). Its main advantage is that the association between each pair of indicators depends on only one parameter that has a straightforward interpretation and is the analogue of the correlation coefficient for the bivariate normal distribution. Further, in the resulting extended IRT model, the interpretation of the latent variables that was suggested at the beginning of Section 2.1 does not change.

For each pair (j_1, j_2) of discretized indicators, the Plackett association parameter corresponds to the following *global log-odds ratio* (Douglas *et al*., 1991):

$$\log\left\{\frac{p(y_{ij_1}<v_1, y_{ij_2}<v_2|u_i)\,p(y_{ij_1}\geqslant v_1, y_{ij_2}\geqslant v_2|u_i)}{p(y_{ij_1}<v_1, y_{ij_2}\geqslant v_2|u_i)\,p(y_{ij_1}\geqslant v_1, y_{ij_2}< v_2|u_i)}\right\}=\tau_{j_1j_2}, \qquad v_1, v_2=1,\ldots,s-1. \quad (4)$$

Note that expression (4) does not depend on the specific categories v_1 and v_2 but only on the pair of indicators that are involved. When $\tau_{j_1j_2}$ is equal to 0, we have conditional independence between the pair of indicators given the latent variable u_i, so local independence holds when $\tau_{j_1j_2}=0$ for all possible pairs (j_1, j_2). Moreover, $\tau_{j_1j_2}>0$ corresponds to a positive association (as the first indicator increases, the second also tends to increase), whereas $\tau_{j_1j_2}<0$ corresponds to a negative association. Testing the null hypothesis that $\tau_{j_1j_2}=0$ against the alternative that $\tau_{j_1j_2}\neq 0$ may be performed by using a standard likelihood ratio test, whose test statistic has an asymptotic $\chi^2(1)$ distribution under the null hypothesis. It is clear that the result of this test, as well as the results of testing all the other assumptions that were discussed in Section 2.1, may depend on the chosen number of classes k. Moreover, relative to expression (3), the number of additional parameters that are required by this extension is equal to the number of pairs of responses for which the association parameter defined in expression (4) is not constrained to be 0.

Another assumption that may be problematic is unidimensionality, which has already been discussed at the beginning of Section 2.1. This assumption may be tested against multi-dimensionality (see Bartolucci (2007), among others). In particular, given the number k of support points, we can use a likelihood ratio test to compare the unidimensional model against a multi-dimensional model. The multi-dimensional model may be formulated by assuming that for each journal i there is a latent vector $\mathbf{u}_i=(u_{i1},\ldots,u_{it})$ of $t>1$ (instead of a single latent variable u_i) such that

$$\log\left\{\frac{p(y_{ij}\geqslant v|\mathbf{u}_i)}{p(y_{ij}<v|\mathbf{u}_i)}\right\}=\alpha_j(u_{id_j}-\beta_{jv}), \qquad v=1,\ldots,s-1,$$

where the index d_j denotes the latent variable measured by indicator j and ranges from 1 to t. The latent vector $\mathbf{u}_i$ is still assumed to have a discrete distribution with support points $\boldsymbol{\xi}_1,\ldots,\boldsymbol{\xi}_k$ and mass probabilities $\pi_1,\ldots,\pi_k$. Considering the identifiability constraints, the number of additional free parameters that are required by this extension is equal to $(k-2)(t-1)$, which therefore makes sense only when $k>2$. Further details on this extension may be found in Bacci *et al.* (2014).

Finally, the assumption that all latent variables u_i (or vectors of latent variables $\mathbf{u}_i$ in the multi-dimensional case) have the same distribution may be relaxed by introducing a vector of covariates $\mathbf{w}_i$, for $i=1,\ldots,n$. In the unidimensional case, there is no loss of generality in reordering the support points of the latent distribution so that $\xi_1<\ldots<\xi_k$. We then assume a parameterization of the conditional distribution of each latent variable u_i given $\mathbf{w}_i$ based on cumulative logits:

$$\log\left\{\frac{p(u_i>\xi_h|\mathbf{w}_i)}{p(u_i\leqslant\xi_h|\mathbf{w}_i)}\right\}=\log\left(\frac{\pi_{h+1,i}+\ldots+\pi_{ki}}{\pi_{1i}+\ldots+\pi_{hi}}\right)=\phi_h+\mathbf{w}_i'\boldsymbol{\psi},\qquad h=1,\ldots,k-1,\tag{5}$$

where $\pi_{hi}=p(u_i=\xi_h|\mathbf{w}_i)$, $\phi_1,\ldots,\phi_{k-1}$ are ordered intercepts and $\boldsymbol{\psi}$ is a vector of regression coefficients. In practice, this is a proportional odds model (McCullagh, 1980) for the probability of belonging to the different latent classes. Unlike the multinomial logit parameterization, the vector $\boldsymbol{\psi}$ is the same for each latent class h, making the interpretation easier. Moreover, the additional number of free parameters due to the inclusion of covariates in expression (5) is simply equal to the dimension of the vector $\mathbf{w}_i$.

2.3. Manifest and posterior distributions

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. Its finite mixture nature derives from considering the distribution of the latent variables as discrete.

Under the basic assumptions 1–3 in Section 2.1, the *manifest distribution* of $\mathbf{y}_i$ may be expressed as

$$p(\mathbf{y}_i)=\sum_{h=1}^{k}\pi_h\prod_{j=1}^{r}\lambda_{j|h}(y_{ij}).\tag{6}$$

This manifest distribution is key for maximum likelihood estimation of the model parameters. The *posterior distribution* of the latent variable u_i, namely its conditional distribution given the vector $\mathbf{y}_i$ of observed indicators, has probability mass function

$$p(u_i=\xi_h|\mathbf{y}_i)=\frac{\pi_h\prod_{j=1}^{r}\lambda_{j|h}(y_{ij})}{p(\mathbf{y}_i)}.\tag{7}$$

This is used to assign every sample unit (i.e. journal) to a given latent class or cluster. Specifically, once the model has been estimated, unit i is assigned to latent class h if

$$h=\underset{g=1,\ldots,k}{\arg\max}\; p(u_i=\xi_g|\,\mathbf{y}_i).\tag{8}$$

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

$$\hat{u}_i=\sum_{h=1}^{k}\xi_h\,p(u_i=\xi_h|\mathbf{y}_i).\tag{9}$$

It is worth noting that posterior probabilities also provide a measure of the classification error when a journal is assigned to a particular latent class by using rule (8). In particular, the amount of classification error for journal i may be directly measured by the distance of $\max_{g=1,\ldots,k} p(u_i = \xi_g|\mathbf{y}_i)$ from 1. Moreover, even if the latent journal impact is modelled as discrete, its predicted value computed by equation (9) ranges on a continuous scale. To clarify this point, let i be a journal for which we do not have a 'neat' assignment, i.e. for this journal there are two latent classes, say the first two, with posterior probabilities that are significantly greater than 0. Then, the predicted value $\hat{u}_i$ will be intermediate between the support points ξ_1 and ξ_2, closer to ξ_1 when $p(u_i = \xi_1|\mathbf{y}_i) > p(u_i = \xi_2|\mathbf{y}_i)$ and closer to ξ_2 when the opposite is true. This is of particular interest when we regard the discrete distribution assumed for the latent variables as an approximation to a continuous distribution.

When there are missing data, we compute the manifest distribution of the discretized indicators as

$$p(\mathbf{y}_i) = \sum_{h=1}^{k} \pi_h \prod_{\substack{j=1 \\ m_{ij}=0}}^{r} \lambda_{j|h}(y_{ij}), \tag{10}$$

where m_{ij} is a binary indicator equal to 1 if y_{ij} is missing and to 0 if it is observed. This amounts to assuming that the data are missing at random in the sense of Little and Rubin (2002). In our context, missingness at random 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, and the available covariates. We consider this assumption to be sufficiently realistic because, as discussed in Section 4, missing values of certain indicators tend to be observed for journals with a lower level of the h-index or for specific covariate values.

The manifest distribution (10) may be simply modified to take into account failure of some of the model assumptions, as discussed in Section 2.2. For example, when we allow for conditional dependence between a pair (j_1, j_2) of indicators, the manifest distribution becomes

$$p(\mathbf{y}_i) = \sum_{h=1}^{k} \pi_h \lambda_{j_1 j_2|h}(y_{ij_1}, y_{ij_2}) \prod_{\substack{j=1 \\ j\neq j_1, j\neq j_2}}^{r} \lambda_{j|h}(y_{ij}), \tag{11}$$

where $\lambda_{j_1 j_2|h}(y_{ij_1}, y_{ij_2})$ refers to the conditional (bivariate) distribution of (y_{ij_1}, y_{ij_2}) given $u_i = \xi_h$. This distribution depends on the conditional (univariate) probabilities $\lambda_{j_1|h}(y_{ij_1})$ and $\lambda_{j_2|h}(y_{ij_2})$ and the association parameter $\tau_{j_1 j_2}$ that is defined in expression (4). A similar extension is available for the posterior distribution of u_i. More complex expressions arise when more than one pair of responses are allowed to be conditionally associated, in which case we may use the rules in Colombi and Forcina (2001), or when some indicators have missing values.

Finally, in the multi-dimensional case the above expressions for the manifest and posterior distributions remain unchanged, because we essentially have the same finite mixture model, except that a vector of supports points $\boldsymbol{\xi}_h$ is now associated with each latent class $h = 1, \ldots, k$. Moreover, when the distribution of the latent variables depends on covariates, we only need to replace the probabilities π_h with unit-specific probabilities π_{hi}. Given the parameters ϕ_h and $\boldsymbol{\psi}$, these probabilities are obtained by simply inverting expression (5), and the manifest and posterior distributions are denoted by $p(\mathbf{y}_i|\mathbf{w}_i)$ and $p(u_i|\mathbf{y}_i, \mathbf{w}_i)$ respectively.

3. Likelihood inference

Given n vectors of discrete indicators $\mathbf{y}_1, \ldots, \mathbf{y}_n$, one for each journal, inference is based on the sample log-likelihood

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

where $\boldsymbol{\theta}$ is the vector containing all the model parameters and $p(\mathbf{y}_i)$ is the manifest probability of $\mathbf{y}_i$, which depends on $\boldsymbol{\theta}$ and is computed according to expression (6), or its extended version which depends on the specific model formulation. When unit-specific covariates collected in the vector $\mathbf{w}_i$ are available, we substitute $p(\mathbf{y}_i)$ with $p(\mathbf{y}_i|\mathbf{w}_i)$ in the above expression for $l(\boldsymbol{\theta})$.

To maximize $l(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, we use the version of the EM algorithm (Dempster *et al.*, 1977) that was described in Bacci *et al.* (2014), to which we refer for details. This implementation is available in the R package `MultiLCIRT` (available from `http://CRAN.R-project.org/package=MultiLCIRT`) that, in its current version, also addresses multi-dimensionality and allows for covariates. We also implemented an R extension of some of the functions that are included in this package to deal with the model that allows for local dependence; the extended functions are available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

In the remainder of this section we first provide a brief description of the EM algorithm. Then, we address the problem of model selection.

3.1. Estimation algorithm

The EM algorithm is based on the *complete-data* log-likelihood. For the baseline IRT model that was described in Section 2.1, this log-likelihood is equal to

$$l^*(\boldsymbol{\theta})=\sum_{h=1}^{k}\sum_{i=1}^{n}z_{hi}\log\left\{\pi_h\prod_{j=1}^{r}\lambda_{j|h}(y_{ij})\right\},\tag{12}$$

where z_{hi} is an (unobserved) binary indicator equal to 1 if $u_i=\xi_h$ and to 0 otherwise. Slightly more complex versions of this expression are used for the complete-data log-likelihood of the extended IRT models that were described in Section 2.2, taking into account the results that were derived in Section 2.3.

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

(a) *E-step*: compute the conditional expected value of $l^*(\boldsymbol{\theta})$ given the observed data and the current value of the parameters.
(b) *M-step*: maximize the above expected value with respect to $\boldsymbol{\theta}$ to obtain an updated estimate of the parameter vector.

The E-step computes the expected value of z_{hi} given $\mathbf{y}_i$ (and possibly $\mathbf{w}_i$) for every h and i through the posterior probabilities, defined in expression (7) for the initial model, and then substitutes these expected values in equation (12). At the M-step, the resulting function is maximized with respect to $\boldsymbol{\theta}$. When the distribution of the latent variable is the same for $i=1,\ldots,n$, the existence of a closed form solution for the probabilities π_h makes the maximization problem easier, whereas updating the other parameters requires only simple iterative algorithms. In contrast, with unit-specific covariates we need an iterative algorithm also to update the parameters in expression (5).

A crucial point is the initialization of the EM algorithm, as the likelihood function may present several local maxima. For the IRT model that was defined in Section 2.1, we rely on two different types of initialization: deterministic and random, both based on guessing initial values of the class weights π_h and the conditional response probabilities $\lambda_{j|h}(v)$. In the deterministic initialization, we set $\pi_h=1/k$, $h=1,\ldots,k$, and choose $\lambda_{j|h}(v)$ so that, for $j=1,\ldots,r$, these probabilities are increasingly ordered in h for $v>(s-1)/2$ and are decreasingly ordered in h

for $v < (s-1)/2$, with $h=1,\ldots,k$ and $v=0,\ldots,s-1$. This rule guarantees that the adding-up constraint $\Sigma_{v=0}^{s-1}\lambda_{j|h}(v)=1$ is satisfied for all h and j. In the random initialization, first we draw every π_h and $\lambda_{j|h}(v)$ from a uniform distribution between 0 and 1; then we normalize them in a suitable way. Similar rules are implemented when an extended model, formulated as described in Section 2.2, is adopted.

3.2. Model selection

To apply the ranking strategy proposed, we must select the number k of support points (or latent classes) of the distribution of u_i. Of course, this is not needed when this number is fixed *a priori*. Among the model selection criteria that are available in the literature on finite mixture and latent class models (McLachlan and Peel (2000), chapter 6), we suggest the Bayesian information criterion (BIC) of Schwarz (1978). This criterion is based on minimization of the index

$$\mathrm{BIC} = -2\,l(\hat{\boldsymbol{\theta}}) + \log(n)\#\mathrm{par}, \tag{13}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under the model of interest and #par is the associated number of parameters, which is defined in equation (3) for the baseline IRT model. The BIC aims at selecting the model that is the best compromise between goodness of fit (measured by the log-likelihood) and complexity (measured by the number of parameters).

In practice, the BIC is similar to the Akaike information criterion (AIC) of Akaike (1973), which is based on the index

$$\mathrm{AIC} = -2\,l(\hat{\boldsymbol{\theta}}) + 2\,\#\mathrm{par}. \tag{14}$$

The main difference of the AIC with respect to the BIC is the smaller penalty for lack of parsimony when $\log(n) > 2$. Consequently, the AIC tends to select larger models with respect to the BIC, and for this reason the latter is usually preferred. Moreover, many simulation studies (e.g. Dias (2006)) show that the BIC tends to outperform the AIC in selecting the correct number of classes of a latent class model.

A different principle is behind another selection criterion, the normalized entropy criterion (NEC) that was proposed by Celeux and Soromenho (1996), which is specifically applied to the choice of the number of latent classes. This criterion takes into account the separation between the latent classes or, equivalently, the quality of the classification. For $k \geqslant 2$, the NEC is based on minimization of the index

$$\mathrm{NEC} = -\frac{\sum_{i=1}^{n}\sum_{h=1}^{k}\hat{p}(u_i=\xi_h|\mathbf{y}_i)\log\{\hat{p}(u_i=\xi_h|\mathbf{y}_i)\}}{l(\hat{\boldsymbol{\theta}}) - \hat{l}_1}, \tag{15}$$

where the numerator is an entropy measure based on the posterior probabilities of the latent classes and $\hat{l}_1$ is the log-likelihood of the model with only one latent class, i.e. the independence model. For $k=1$, we conventionally set NEC$=1$; see also Biernacki *et al.* (1999). The use of the NEC typically leads to selecting a very small number of latent classes that are strongly separated.

Finally, in selecting a suitable model for the analysis of the available data, we suggest beginning from the choice of the number of latent classes, k, under the initial assumptions that were described in Section 2.1. Then, using the same model selection criterion, say BIC, we suggest selecting some of the possible extensions that were described in Section 2.2.

4. Empirical application

We illustrate our approach by considering the problem that is faced in the Italian VQR of ranking the journals that are considered relevant for a given scientific area. In most areas covered by the VQR, journal papers were the main type of research products submitted for evaluation. For these areas, an important input to the evaluation process was the ranking of the journals where these papers were published.

4.1. *The Italian research assessment exercise*

The VQR involved all state universities, all private universities granting publicly recognized academic degrees and several public research centres. Researchers who were affiliated to these institutions were asked to submit for evaluation various research products (i.e. journal papers, books, book chapters and patents) produced during the period 2004–2010. The typical number of products submitted by each researcher was six for public research centres and three for universities (under the convention that half of the time of the academic staff is devoted to teaching). The evaluation process assigned each product to one of six merit classes, namely 'excellent' (corresponding to 'the highest 20% of the quality ranking shared by the international scientific community'), 'good' (corresponding to the top 20–40% segment), 'acceptable' (corresponding to the top 40–50% segment), 'limited' (corresponding to the lowest 50%), 'not assessable' and 'fraud or plagiarism'. On the basis of its classification, each product received a numerical score equal to 1 for excellent, 0.80 for good, 0.50 for acceptable, 0 for limited, −1 for not assessable and −2 for fraud or plagiarism. Finally, participating institutions (and their departments) were ranked on the basis of average score received by the products that were submitted by their researchers. Individual researchers were not ranked.

In our application, we focus on a subset of the journals that were considered relevant for the area economics and statistics, whose products were evaluated by one of the groups of experts formed by Agenzia di Valutazione del Sistema Universitario e della Ricerca: GEV 13 as the area economics and statistics is 13 in the Italian classification of scientific areas. Journal papers in this area were evaluated via a bibliometric analysis based on a preliminary classification of the journals where they had been published and the number of citations that they had received. For a random sample of journal papers, the bibliometric analysis was complemented by informed peer review. (We refer to Bertocchi *et al.* (2014) for more detail on the procedures that were adopted by GEV 13.) To avoid different rankings across subareas, GEV 13 associated each journal with one and only one of its four subareas: business, management and finance; economics; economic history and history of economic thought; statistics and financial mathematics. For simplicity we focus on the set of $n=445$ journals that were associated with the subarea statistics and financial mathematics.

Some of the journals are listed in Tables 1 and 2, along with the four bibliometric indicators that we used. The full data set is available from `http://www.anvur.org/index.php?option=com_content&view=article&id=92&Itemid=390&lang=it`.

4.2. *Data*

To be as close as possible to the actual problems that are faced in the VQR, we base our ranking on exactly the same bibliometric indicators as used by GEV 13, namely IF, IF5, AIS and the h-index.

GEV 13 obtained IF, IF5 and AIS from ISI–Thomson Reuters with reference to the year 2010 and collected the h-index from Google Scholar in April 2012 with reference to the period 2004–2010. Whereas IF, IF5 and AIS have missing values for some journals, the h-index is available for all journals considered by GEV 13. Note that IF, IF5 and AIS are based on the particular selection of journals by one specific commercial supplier and measure visibility in recent years (2 years for IF and 5 years for IF5 and AIS), not lifetime impact. In addition,

Table 1. List of journals assigned to the third class under model1†

Title	*IF*	*IF5*	*AIS*	*h-index*	*post3*	*post4*	*Predicted value*	
							Model 1	*Model 2*
Advances in Applied Probability	0.720	0.967	0.980	29	0.9591	0.0362	0.9036	0.4352
Algorithmica	1.239	1.227	0.721	34	0.9157	0.0828	0.9398	0.6275
American Mathematical Monthly	—‡	—‡	—‡	19	0.4871	0.1026	0.5500	0.3358
Annales de l'Institut Henri Poincaré—Probabilités et Statistiques	0.759	0.900	1.083	18	0.5712	0.0007	0.4642	0.2881
Annals of Operations Research	0.675	1.223	0.741	48	0.9641	0.0301	0.8981	0.6283
Applied Mathematics & Optimization	0.881	1.061	0.822	20	0.9188	0.0034	0.8090	0.6270
Applied Mathematics Letters	1.155	1.127	0.481	41	0.8649	0.0118	0.7704	0.5974
Applied Psychological Measurement	1.137	1.497	0.854	24	0.9360	0.0630	0.9263	0.6265
Biometrical Journal	1.438	1.273	0.822	24	0.9392	0.0284	0.8711	0.6278
British Journal of Mathematical and Statistical Psychology	1.419	1.413	0.920	21	0.9304	0.0691	0.9309	0.6254
Canadian Journal of Statistics–Revue Canadienne de Statistique	0.689	1.175	1.163	20	0.5395	0.0007	0.4331	0.6217
Chaos, Solitons & Fractals	1.268	1.729	0.538	83	0.8984	0.1009	0.9533	0.7165
Combinatorics Probability & Computing	0.990	1.008	1.465	28	0.7059	0.2935	1.0894	0.6305
Complexity	1.367	1.190	0.528	43	0.9668	0.0165	0.8780	0.6260
Computational Optimization and Applications	1.274	1.470	0.916	34	0.6216	0.3784	1.1501	0.6365
Computational Statistics & Data Analysis	1.089	1.363	0.754	55	0.5562	0.4437	1.1963	0.6378
Demographic Research	1.531	1.582	0.708	38	0.8730	0.1267	0.9718	0.6722
Differential and Integral Equations	—‡	—‡	—‡	18	0.4595	0.0770	0.4792	0.3070
Discrete Event Dynamic Systems	0.872	1.011	0.719	24	0.5196	0.0004	0.4132	0.6020
Econometric Theory	1.015	1.264	1.541	40	0.7217	0.2776	1.0781	0.6877
Econometrics Journal	0.691	1.166	1.253	30	0.8656	0.1312	0.9722	0.6540
Educational and Psychological Measurement	0.831	1.434	0.658	34	0.8402	0.1573	0.9914	0.6269
Electronic Journal of Probability	0.946	1.044	1.343	13	0.5582	0.0052	0.4592	0.6085
Electronic Journal of Statistics	1.025	1.208	1.411	21	0.9074	0.0150	0.8173	0.6221
Environmental and Ecological Statistics	1.645	1.641	0.872	22	0.9372	0.0622	0.9260	0.6486
European Journal of Ageing: Social, Behavioural and Health Perspectives	1.119	—‡	—‡	24	0.7299	0.0380	0.6822	0.5937
European Journal of Population	1.049	1.966	1.088	27	0.6473	0.3518	1.1303	0.7252
Extremes	1.053	—‡	—‡	17	0.7830	0.0562	0.7650	0.4911
IEEE Transactions on Reliability	1.288	1.698	0.701	38	0.8012	0.1985	1.0227	0.7716
IIE Transactions	1.186	1.535	0.777	42	0.6875	0.3125	1.1034	0.8038
IMA Journal of Management Mathematics	0.608	1.980	1.006	16	0.7544	0.2377	1.0430	0.7232
Informs Journal on Computing	1.172	1.450	0.903	34	0.5562	0.4437	1.1963	0.6378
Insurance: Mathematics and Economics	1.178	1.451	0.780	41	0.6397	0.3603	1.1373	0.6351
International Migration	0.556	1.047	0.520	33	0.6853	0.0024	0.5786	0.6130
International Statistical Review	0.860	0.852	0.625	20	0.5492	0.0004	0.4422	0.3074
Journal of Agricultural, Biological and Environmental Statistics	0.722	1.220	0.744	20	0.9299	0.0044	0.8214	0.6173
Journal of Applied Probability	0.768	0.866	0.767	29	0.9591	0.0362	0.9036	0.3814
Journal of Biopharmaceutical Statistics	1.073	1.285	0.602	26	0.9769	0.0124	0.8809	0.6001
Journal of Biosocial Science	1.217	1.330	0.454	25	0.8981	0.0078	0.7961	0.5070

(*continued*)

Table 1 (*continued*)

Title	*IF*	*IF5*	*AIS*	*h-index*	*post3*	*post4*	*Predicted value*	
							Model 1	*Model 2*
Journal of Chemometrics	1.377	1.858	0.539	29	0.8773	0.1221	0.9684	0.6416
Journal of Classification	1.100	1.225	0.704	18	0.7908	0.0015	0.6804	0.5061
Journal of Computational and Applied Mathematics	1.030	1.299	0.571	51	0.8556	0.1440	0.9840	0.6261
Journal of Empirical Finance	0.807	—‡	—‡	46	0.7506	0.1375	0.8702	0.6368
Journal of Ethnic and Migration Studies	1.041	1.424	0.646	45	0.8773	0.1221	0.9684	0.6239
Journal of Financial Econometrics	0.846	—‡	—‡	36	0.7148	0.2298	0.9909	0.6808
Journal of Forecasting	0.655	0.866	0.525	28	0.9446	0.0073	0.8407	0.3219
Journal of Global Optimization	1.160	1.433	0.698	38	0.8688	0.1307	0.9746	0.6260
Journal of Mathematical Analysis and Applications	1.174	1.345	0.718	66	0.8556	0.1440	0.9840	0.6261
Journal of Mathematical Psychology	1.582	1.833	1.017	27	0.9304	0.0691	0.9309	0.6853
Journal of Mathematical Sciences	—‡	—‡	—‡	23	0.4595	0.0770	0.4792	0.5834
Journal of Multivariate Analysis	1.010	1.180	0.917	37	0.9380	0.0599	0.9229	0.6351
Journal of the Operational Research Society	1.102	1.481	0.603	46	0.8556	0.1440	0.9840	0.6261
Journal of Optimization Theory and Applications	1.011	1.209	0.706	36	0.9623	0.0270	0.8913	0.6294
Journal of Quality Technology	1.377	2.132	1.026	26	0.5081	0.4914	1.2295	0.9763
Journal of the Royal Statistical Society Series C—Applied Statistics	0.645	1.284	0.961	29	0.5689	0.4309	1.1870	0.6283
Journal of Time Series Analysis	0.678	0.888	0.871	25	0.9188	0.0034	0.8090	0.4481
Lifetime Data Analysis	0.873	1.014	0.857	18	0.8982	0.0026	0.7873	0.5917
Networks	0.991	1.167	0.728	32	0.9157	0.0828	0.9398	0.6283
Networks and Heterogeneous Media	0.909	1.381	0.836	17	0.9406	0.0253	0.8672	0.5422
Nonlinear Analysis, Theory, Methods & Applications	1.279	1.409	0.565	44	0.8773	0.1221	0.9684	0.6239
Operations Research Letters	0.743	1.007	0.661	34	0.9446	0.0073	0.8407	0.6283
Operations Research—Spektrum	2.030	1.864	0.714	33	0.8474	0.1523	0.9901	0.6772
Pharmaceutical Statistics	1.630	1.467	0.729	19	0.9394	0.0599	0.9243	0.4037
Population Bulletin	1.182	1.741	1.007	6	0.9381	0.0426	0.8938	0.6990
Population Studies	0.974	1.723	0.902	28	0.5949	0.4051	1.1689	0.7101
Population, Space and Place	1.429	1.500	0.581	30	0.8730	0.1267	0.9718	0.7165
Probabilistic Engineering Mechanics	1.252	1.306	0.721	28	0.6397	0.3603	1.1373	0.5938
Probability in the Engineering and Informational Sciences	0.971	0.966	0.754	20	0.9521	0.0051	0.8444	0.3669
Quantitative Finance	0.590	0.968	0.687	38	0.9276	0.0056	0.8212	0.3806
Queueing Systems	0.802	1.207	0.855	28	0.9538	0.0421	0.9084	0.6270
Set-valued Analysis	1.418	1.120	0.860	19	0.9147	0.0126	0.8204	0.5736
Social Indicators Research	1.000	1.239	0.409	48	0.8240	0.0080	0.7238	0.5838
Statistical Modelling	0.714	1.021	0.756	20	0.8941	0.0024	0.7832	0.6271
Statistica Sinica	0.956	1.020	0.969	32	0.9549	0.0276	0.8851	0.6274
Stochastic Environmental Research and Risk Assessment	1.777	1.700	0.460	22	0.9279	0.0086	0.8266	0.6274
Studies in Family Planning	1.778	1.818	0.698	27	0.9702	0.0262	0.8976	0.6401
System Dynamics Review	0.667	1.586	0.506	26	0.9211	0.0120	0.8258	0.6445
Systems & Control Letters	1.412	1.768	1.022	54	0.5789	0.4211	1.1803	0.9193
Test	1.036	1.108	1.176	18	0.7323	0.0018	0.6237	0.5373

†post3 and post4 stand for the posterior probability that a certain journal is in the third and in the fourth class respectively.
‡Not applicable.

Table 2. List of journals assigned to the fourth class under model 1†

Title	*IF*	*IF5*	*AIS*	*h-index*	*post3*	*post4*	*Predicted value*	
							Model 1	*Model 2*
Ageing and Society	1.309	1.900	0.559	39	0.3858	0.6140	1.3164	0.7253
American Statistician	0.981	1.322	0.924	29	0.4994	0.5006	1.2364	0.6282
Annals of Applied Probability	1.120	1.447	1.595	44	0.2143	0.7857	1.4380	0.6972
Annals of Applied Statistics	1.746	2.443	2.072	28	0.0211	0.9789	1.5746	0.9896
Annals of Probability	1.470	1.665	1.996	45	0.1388	0.8612	1.4914	0.9826
Annals of Statistics	2.940	3.274	3.260	76	0.0061	0.9939	1.5851	1.0016
Bayesian Analysis	1.213	2.756	2.237	25	0.3744	0.6245	1.3230	0.9813
Bernoulli	1.000	1.284	1.577	33	0.2675	0.7325	1.4003	0.6817
Biometrics	1.764	2.204	1.594	47	0.0090	0.9910	1.5831	1.0014
Biometrika	1.833	2.352	2.393	47	0.0090	0.9910	1.5831	1.0014
Biostatistics	2.769	3.303	2.312	45	0.0134	0.9866	1.5800	1.0012
Chaos, an Interdisciplinary Journal of Nonlinear Science	2.081	2.134	1.082	37	0.0576	0.9424	1.5487	0.9997
Chemometrics and Intelligent Laboratory Systems,	2.222	2.415	0.645	44	0.2109	0.7891	1.4404	0.9752
Computational Biology and Bioinformatics, IEEE, ACM Transactions	1.664	2.171	0.950	33	0.0598	0.9402	1.5472	0.9924
Computers & Operations Research	1.769	2.250	0.886	75	0.0620	0.9380	1.5457	0.9923
Decision Support Systems	2.135	2.573	0.708	68	0.2109	0.7891	1.4404	0.9752
Demography	2.465	3.817	2.207	57	0.0061	0.9939	1.5851	1.0016
Ecological Modelling	1.769	2.439	0.753	61	0.0620	0.9380	1.5457	0.9923
Econometric Reviews	1.088	1.400	1.346	31	0.1722	0.8278	1.4677	0.6553
European Journal of Operational Research	2.159	2.513	0.886	108	0.0620	0.9380	1.5457	0.9923
Finance and Stochastics	1.326	1.870	2.016	43	0.0310	0.9690	1.5676	0.9717
Fuzzy Sets and Systems	1.875	2.250	0.591	67	0.2109	0.7891	1.4404	0.9752
Games and Economic Behavior	1.017	1.503	1.817	63	0.1722	0.8278	1.4677	0.9826
International Family Planning Perspectives	2.118	2.575	0.902	30	0.0446	0.9554	1.5580	0.9951
International Migration Review	1.188	2.145	1.236	47	0.0519	0.9481	1.5528	0.9931
Journal of the American Statistical Association	2.063	3.439	3.280	73	0.0061	0.9939	1.5851	1.0016
Journal of Applied Econometrics	1.341	2.268	2.172	55	0.0179	0.9821	1.5768	1.0011
Journal of Business & Economic Statistics	1.693	2.433	2.804	50	0.0077	0.9923	1.5840	1.0015
Journal of Computational Biology	1.600	2.033	0.907	45	0.0576	0.9424	1.5487	0.9939
Journal of Computational and Graphical Statistics	1.206	1.848	1.576	39	0.2143	0.7857	1.4380	0.9793
Journal of Econometrics	1.815	2.823	3.016	91	0.0108	0.9892	1.5818	1.0013
Journal of Educational and Behavioral Statistics	1.644	2.474	1.862	27	0.0840	0.9160	1.5301	0.9965
Journal of Risk and Uncertainty	1.558	1.953	1.317	36	0.0134	0.9866	1.5800	0.8382
Journal of the Royal Statistical Society, Series A—Statistics in Society	2.570	2.527	1.822	40	0.0090	0.9910	1.5831	1.0014
Journal of the Royal Statistical Society, Series B—Statistical Methodology	3.500	5.086	4.822	53	0.0090	0.9910	1.5831	1.0014
Journal of Statistical Mechanics: Theory and Experiment	1.822	2.169	1.088	39	0.0778	0.9222	1.5345	0.9984
Journal of Statistical Physics	1.447	1.534	0.950	47	0.4358	0.5642	1.2814	0.8683
Journal of Statistical Software	2.647	3.654	1.735	36	0.0100	0.9900	1.5824	1.0014
Mathematical Finance	1.052	1.801	1.892	47	0.2143	0.7857	1.4380	0.9793
Mathematical Programming	1.970	2.781	1.951	55	0.0108	0.9892	1.5818	1.0013
Mathematics of Computation	1.382	1.565	1.276	42	0.1137	0.8863	1.5091	0.9508
Mathematics of Operations Research	1.145	1.478	1.423	34	0.1722	0.8278	1.4677	0.6553
Multivariate Behavioral Research	1.290	3.295	2.062	28	0.0143	0.9857	1.5794	0.9947

(*continued*)

Table 2 (*continued*)

Title	*IF*	*IF5*	*AIS*	*h-index*	*post3*	*post4*	*Predicted value*	
							Model 1	*Model 2*
Nonlinear Analysis Real World Applications	2.138	2.039	0.625	40	0.2488	0.7512	1.4136	0.9718
Operations Research	2.000	2.708	1.928	61	0.0061	0.9939	1.5851	1.0016
Oxford Bulletin of Economics and Statistics	1.182	1.622	1.225	46	0.1960	0.8040	1.4509	0.9309
Pattern Recognition Letters	1.235	1.897	0.696	67	0.3073	0.6926	1.3721	0.7517
Perspectives on Sexual and Reproductive Health	2.075	3.842	1.367	35	0.0104	0.9896	1.5821	0.9993
Population and Development Review	1.507	2.381	1.362	43	0.0077	0.9923	1.5840	1.0001
Probability Theory and Related Fields	1.590	1.625	1.985	40	0.1588	0.8412	1.4772	0.9517
Psychometrika	1.778	1.804	1.181	28	0.4358	0.5642	1.2814	0.7058
Reliability Engineering & System Safety	1.899	2.023	0.660	52	0.2109	0.7891	1.4404	0.9752
Risk Analysis	2.096	2.344	0.788	57	0.0446	0.9554	1.5580	0.9951
Scandinavian Journal of Statistics	0.835	1.326	1.354	28	0.2094	0.7905	1.4414	0.6292
SIAM Journal on Applied Mathematics	1.529	1.824	1.062	37	0.4358	0.5642	1.2814	0.8982
SIAM Journal on Control and Optimization	1.297	1.666	1.270	39	0.1421	0.8579	1.4890	0.9508
SIAM Journal on Imaging Sciences	4.500	4.500	2.661	16	0.2197	0.7801	1.4339	0.9691
SIAM Journal on Mathematical Analysis	1.797	1.744	1.454	36	0.1137	0.8863	1.5091	0.8982
SIAM Journal on Optimization	2.091	2.566	1.686	47	0.0100	0.9900	1.5824	1.0014
Sociological Methods & Research	2.000	2.448	1.722	30	0.0077	0.9923	1.5840	1.0015
Stata Journal	2.000	3.142	1.964	37	0.0104	0.9896	1.5821	1.0013
Statistical Applications in Genetics and Molecular Biology	1.842	2.182	1.100	30	0.0778	0.9222	1.5345	0.9984
Statistical Methods in Medical Research	1.768	2.541	1.535	29	0.0100	0.9900	1.5824	0.9955
Statistical Science	2.480	3.504	3.383	41	0.0077	0.9923	1.5840	1.0015
Statistics and Computing	1.851	2.339	1.838	31	0.0100	0.9900	1.5824	1.0014
Statistics in Medicine	2.328	2.334	1.330	62	0.0077	0.9923	1.5840	1.0001
Stochastic Processes and Their Applications	0.951	1.381	1.368	38	0.2094	0.7905	1.4414	0.6505
Structural Equation Modeling	2.738	5.611	2.633	40	0.0134	0.9866	1.5800	1.0012
Technometrics	1.560	1.985	1.424	34	0.0061	0.9939	1.5851	0.8982

†post3 and post4 stand for the posterior probability that a certain journal is in the third and in the fourth class respectively.

the selection of journals by ISI–Thomson Reuters is not completely transparent and does not follow quality criteria alone. In contrast, the h-index is based on a much larger but also much more heterogeneous set of indexed sources, also including books, book chapters and publicly available conference proceedings. Although the h-index may better measure the impact of a journal, it does favour older journals.

Table 3 shows descriptive statistics for our $r=4$ indicators. Note that IF is available for only 56.2% of the journals and IF5 and AIS for only 47.4%, whereas the h-index is always available. Also note the lower mean h-index for journals with IF missing compared with those with IF5 and AIS missing. All four indicators are available for a subset of 211 journals. For this subset of journals Table 3 shows also the mean, the variance, the index of skewness, the quartiles and the deciles of the available indicators. Fig. 1 presents their distribution and their scatter plot matrix. In general, we note large differences between the distribution of the h-index for ISI–Thomson Reuters and non-ISI–Thomson Reuters journals, suggesting that the h-index, together with the covariates that are included in the data set, may be a good predictor of the probability that another indicator is missing for a certain journal.

Table 3. Descriptive statistics for the indicators observed

	Results for the following indicators:			
	IF	*IF5*	*AIS*	*h-index*
Missing values (%)	43.8	52.6	52.6	0
Mean	1.056	1.472	0.946	19.766
Mean (given IF missing)	—	—	—	8.446
Mean (given IF5 or AIS missing)	0.492	—	—	9.509
Variance	0.418	0.751	0.480	267.585
Skewness index	1.325	1.526	1.938	1.575
Quartiles				
1st	0.586	0.840	0.506	7.0
2nd	0.954	1.284	0.721	14.0
3rd	1.381	1.867	1.203	28.0
Deciles				
1st	0.370	0.590	0.313	4.0
2nd	0.521	0.766	0.454	6.0
3rd	0.643	0.967	0.553	9.0
4th	0.754	1.108	0.660	12.0
5th	0.954	1.284	0.721	14.0
6th	1.088	1.467	0.871	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

As for the covariates, we include the age of each journal (i.e. the number of years since the journal was first published), its language and the country of the publisher. For age, we consider six classes (1–3, 4–6, 7–10, 11–20, 21–40 and 41 or more years), for language we distinguish only between English and other languages, and for the country of the publisher we consider three categories: one for the USA (USA), one for the UK, Germany or the Netherlands (UK, D, NL), and one for all other countries (other). Table 4 shows how the distribution of the four indicators depends on these covariates. As expected, the percentage of missing values of IF, IF5 and AIS is much higher for more recent than for older journals. The mean value of the available indicators does not show a clear pattern but tends to be lower for journals in the age class 4–6 years. In this regard note that IF, IF5 and AIS are available for only two of the 17 journals in the first age category (1–3 years), so, the exceptionally high value of the mean of these indicators is better regarded as an anomaly. The pattern for the h-index shows instead a clear tendency for this indicator to increase with the age of a journal. Another clear difference also emerges in connection with language, as English journals have on average much higher values of all indicators compared with journals in other languages. The percentage of missing values is also much higher for this second category. Similarly, for journals that are published in the USA, the UK, Germany or the Netherlands we observe a smaller percentage of missing values and a higher mean for all indicators compared with journals that are published in other countries. However the comparison between journals that are published in the USA and those published in the UK, Germany or the Netherlands is not so clear, as the first perform better in terms of IF, IF5 and AIS, but perform a little worse in terms of h-index.

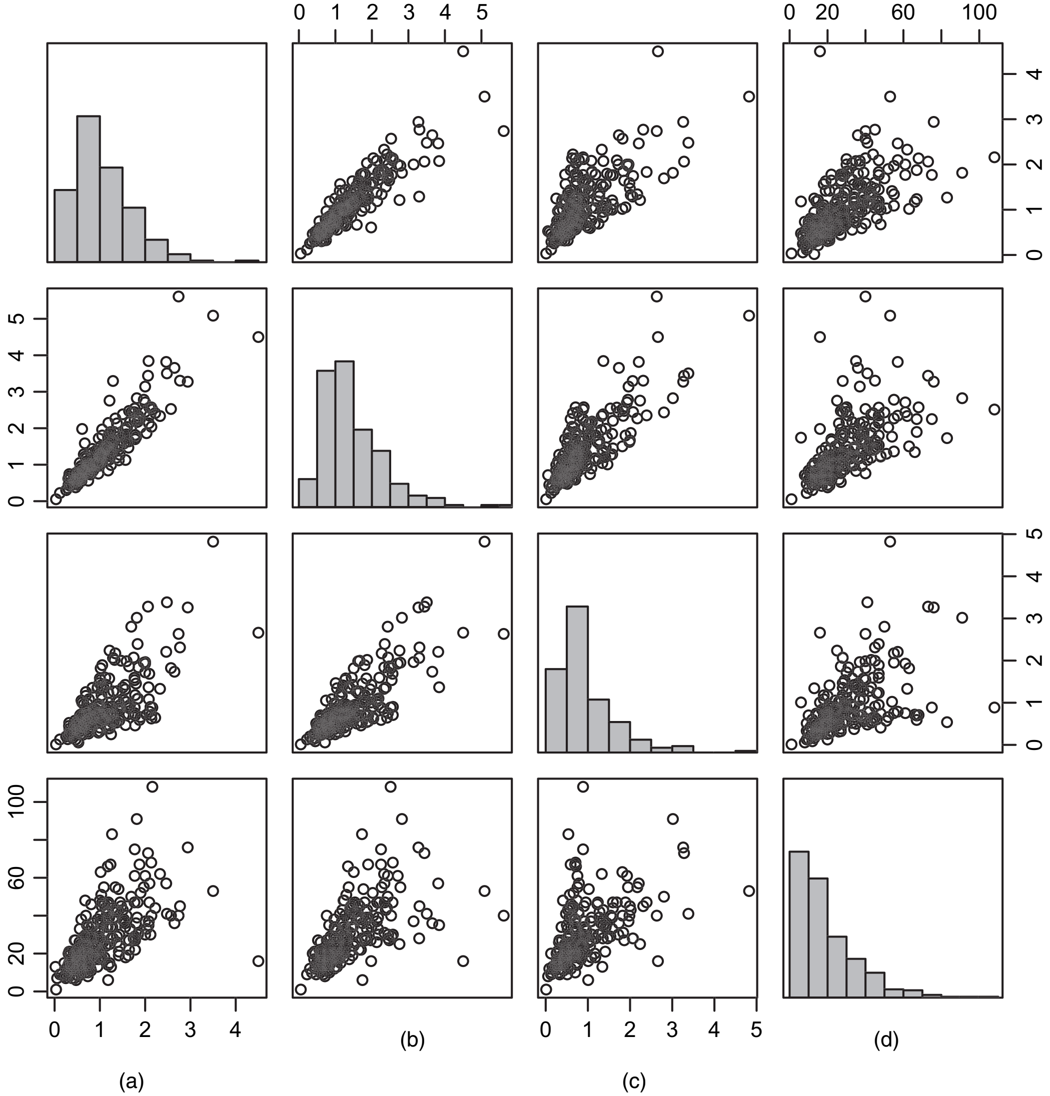

Fig. 1. Scatter plot of the indicators observed: (a) IF; (b) IF5; (c) AIS; (d) *h*-index

4.3. Model fitting

As discussed at the beginning of Section 2, the first step of our strategy consists of discretizing the indicators observed. We present two alternative types of discretization: the first uses as cut-offs the sample quartiles ($s=4$); the second is based on fixed cut-offs giving $s=5$ categories. In the latter case, the cut-offs are (0.50, 0.75, 1.00, 1.50) for IF and AIS, (0.75, 1.00, 1.50, 2.00) for IF5 and (5, 10, 20, 30) for the h-index. To avoid categories containing too few journals, these sets of cut-offs are not equally spaced, i.e. the relationship between the value of the discretized indicators y_{ij} and the cut-offs q_{jv} is not restricted to be linear.

Given the discretized indicators y_{ij}, to select an appropriate model for our data we proceed as discussed in Section 3.2. First we choose the number k of latent classes for the baseline IRT model which assumes local independence, unidimensionality and no covariates. To avoid

Table 4. Descriptive statistics for the indicators observed given the available covariates

Covariate	*Category*	*Number*		*Results for the following indicators:*			
				IF	*IF5*	*AIS*	*h-index*
Age (years)	1–3	17	Missing values (%)	88.2	88.2	88.2	0
			Mean	2.487	2.487	1.586	5.412
	4–6	43	Missing values (%)	74.4	79.1	79.1	0
			Mean	0.791	1.324	0.905	9.605
	7–10	50	Missing values (%)	66.0	78.0	78.0	0
			Mean	1.112	1.731	0.903	13.680
	11–20	112	Missing values (%)	48.2	59.8	59.8	0
			Mean	1.082	1.516	0.991	18.500
	21–40	138	Missing values (%)	26.8	36.2	36.2	0
			Mean	0.982	1.362	0.834	25.428
	$\geqslant 41$	85	Missing values (%)	28.2	34.1	34.1	0
			Mean	1.137	1.549	1.077	23.835
Language	English	430	Missing values (%)	42.8	51.2	51.2	0
			Mean	1.065	1.475	0.949	20.186
	Other	15	Missing values (%)	73.3	93.3	93.3	0
			Mean	0.490	0.992	0.250	7.733
Country	USA	130	Missing values (%)	27.7	31.5	31.5	0
			Mean	1.286	1.694	1.159	25.185
	UK, D, NL	143	Missing values (%)	27.3	36.4	36.4	0
			Mean	1.091	1.483	0.886	27.161
	Other	172	Missing values (%)	69.8	82.0	82.0	0
			Mean	0.568	0.804	0.509	9.523

Table 5. Fit of the baseline IRT models with discretization based on quartiles ($s=4$) and fixed cut-offs ($s=5$) as a function of the number k of latent classes†

k	*Results for quartiles*					*Results for fixed cut-offs*				
	$l(\hat{\theta})$	*#par*	*AIC*	*BIC*	*NEC*	$l(\hat{\theta})$	*#par*	*AIC*	*BIC*	*NEC*
1	−1544.9	12	3113.8	3163.0	1.000	−1761.0	16	3554.0	3619.6	1.000
2	−1343.8	17	2721.5	2791.2	*0.241*	−1568.6	21	3179.2	3265.2	*0.298*
3	−1293.0	19	2624.1	2702.0	0.421	−1482.9	23	3011.8	3106.1	0.366
4	−1273.6	21	2589.1	*2675.2*	0.500	−1457.2	25	2964.5	3066.9	0.480
5	−1271.0	23	*2588.0*	2682.3	0.659	−1447.6	27	*2949.3*	*3059.94*	0.451
6	−1271.0	25	2592.0	2694.5	0.752	−1446.7	29	2951.5	3070.3	0.452

†The smallest values of AIC, BIC and NEC are in italics.

spurious estimates due to multimodality of the sample log-likelihood, we use two types of initialization (deterministic and random) of the EM algorithm. The results, in terms of the AIC, BIC and NEC, are reported in Table 5 for the cases of $s=4$ (quartiles) and $s=5$ (fixed cut-offs). As shown in Table 5, the BIC suggests two models for the data: one with $k=4$ latent classes in the case of quartiles, and one with $k=5$ latent classes in the case of fixed cut-offs. The AIC gives a similar result for the first type of discretization and exactly the same result for the second. The NEC suggests instead only two latent classes for both types of discretization.

Next we select the covariates on the basis of the parameterization (5) for the latent class probabilities and a stepwise forward covariates selection scheme based on the BIC; the results are displayed in Table 6. For both discretizations that were considered (quartiles and fixed cut-offs), the specification preferred includes a journal's age and country, but not its language.

To check the unidimensionality assumption, we look for evidence of bidimensionality. Table 7 shows the values of the BIC for various specifications of bidimensionality. For instance, the first model considered, which is denoted by $\{1\}$,$\{2,3,4\}$, assumes that the first indicator measures a latent trait that is different from that for the other three, whereas the fifth model considered, which is denoted by $\{1,2\}$ $\{3,4\}$, assumes that the first two indicators measure a latent trait that is different from that for the other two. Under both discretizations, the smallest value of the BIC in Table 7 is greater than the smallest value of the BIC in Table 6. We conclude that the assumption of unidimensionality is consistent with the data, so our four indicators represent different measurements of the same latent trait.

Finally, we consider the assumption of local independence by fitting some extended models that include a set of association parameters for specific pairs of responses, as defined in expression (4), under Plackett's formulation of local dependence. The results are presented in Table 8. For the first type of discretization (quartiles), there is evidence of conditional association (given the latent trait) between IF and IF5. In fact, the extended model that includes

Table 6. Allowing for covariates in the model with $k=4$ latent classes and discretization based on quartiles and the model with $k=5$ latent classes and discretization based on fixed cut-offs†

Covariate	*Results for quartiles*			*Results for fixed cut-offs*		
	$l(\hat{\boldsymbol{\theta}})$	*#par*	*BIC*	$l(\hat{\boldsymbol{\theta}})$	*#par*	*BIC*
Age	−1259.0	26	2676.6	−1426.4	32	3047.9
Language	−1269.7	22	2673.5	−1441.9	28	3054.6
Country	−1203.4	23	2547.0	−1375.1	29	2927.0
Age + country	−1180.0	28	*2530.7*	−1345.9	34	*2899.1*
Language + country	−1203.3	24	2552.9	−1374.8	30	2932.5
Age + language + country	−1179.1	29	2535.1	−1344.0	35	2901.4

†The smallest values of the BIC in each column are in italics.

Table 7. Allowing for bidimensionality†

Groups of items	*Results for quartiles*			*Results for fixed cut-offs*		
	$l(\hat{\boldsymbol{\theta}})$	*#par*	*BIC*	$l(\hat{\boldsymbol{\theta}})$	*#par*	*BIC*
$\{1\},\{2,3,4\}$	−1178.7	30	*2540.4*	−1344.9	37	2915.4
$\{2\},\{1,3,4\}$	−1179.2	30	2541.4	−1345.9	37	2917.4
$\{3\},\{1,2,4\}$	−1179.9	30	2542.7	−1345.9	37	2917.4
$\{4\},\{1,2,3\}$	−1179.9	30	2542.7	−1345.0	37	2915.6
$\{1,2\},\{3,4\}$	−1179.7	30	2542.4	−1344.9	37	*2915.4*
$\{1,3\},\{2,4\}$	−1179.2	30	2541.4	−1345.0	37	2915.6

†The smallest values of the BIC in each column are in italics.

Table 8. Inclusion of local dependence between certain pairs of indicators†

Pairs of responses	*Results for quartiles*			*Results for fixed cut-offs*		
	$l(\hat{\boldsymbol{\theta}})$	#*par*	*BIC*	$l(\hat{\boldsymbol{\theta}})$	#*par*	*BIC*
{1,2}	−1161.6	29	*2500.1*	−1330.6	35	2874.6
{1,3}	−1179.9	29	2536.7	−1344.6	35	2902.5
{1,4}	−1179.7	29	2536.3	−1345.8	35	2905.0
{2,3}	−1176.8	29	2530.4	−1344.4	35	2902.2
{2,4}	−1179.3	29	2535.5	−1345.5	35	2904.5
{3,4}	−1179.8	29	2536.5	−1345.6	35	2904.6
{1,2},{1,3}	−1161.1	30	2505.1	−1330.3	36	2880.1
{1,2},{1,4}	−1161.1	30	2505.2	−1329.3	36	2878.2
{1,2},{2,3}	−1160.0	30	2503.0	−1325.1	36	*2869.7*
{1,2},{2,4}	−1160.9	30	2504.7	−1330.1	36	2879.7
{1,2},{3,4}	−1160.1	30	2503.2	−1328.8	36	2877.1

†The smallest values of the BIC in each column are in italics.

this association has a lower BIC (namely 2500.1) compared with the model with local independence, and no association between other pairs of indicators needs to be considered. In contrast, for the second type of discretization (fixed cut-offs), there is evidence of conditional association both between IF and IF5 and between IF5 and AIS. However, although the larger model including these two associations and the model including only the association between IF and IF5 have a very similar BIC, the former seems to have some numerical instability in the estimation.

In conclusion, although the two types of discretization imply some differences in terms of local independence and some changes in the ranking of journals in terms of predicted impact, the IRT models that were selected by the procedure that was suggested in Section 3.2 are based on exactly the same features and differ only in the number of latent classes, namely $k=4$ for the quartiles case and $k=5$ for the fixed cut-offs case. The features that are common to both models are unidimensionality, local dependence only between IF and IF5, and latent distribution affected by two covariates: age of the journal and country of the publisher.

4.4. Latent distributions, clustering and prediction of journal impact

The maximum likelihood estimates that were obtained for the parameters of the selected models under the two different discretizations are presented in Tables 9–11. In particular, Table 9 shows the estimated distribution of the latent variable (support points and probabilities). Note that the support points correspond to classes of journals characterized by increasing impact, whereas the mass probabilities are averaged over all the sample units. Also note that we use the standardization of the latent variables as an identifiability constraint. The first discretization (quartiles) selects four classes of journals, the class with highest impact being the smallest (it includes 15.8% of the journals) and the class of journals with lowest impact being the largest (it includes 33.6% of the journals). The second discretization (fixed cut-offs) selects five classes, with the added class consisting of a relatively small number of journals with very low impact (9.8% of the total) and the top two classes having size and support that are comparable with the top two classes estimated under the first model.

Table 10 reports the estimated intercepts and regression coefficients for the covariates included, namely the age of the journal (the reference category is aged 41 or more years) and the

Table 9. Estimated distribution of the latent variable under the models selected

h	*Results for quartiles*		*Results for fixed cut-offs*	
	$\hat{\xi}_h$	$\hat{\pi}_h$	$\hat{\xi}_h$	$\hat{\pi}_h$
1	−1.158	0.336	−2.763	0.098
2	−0.096	0.317	−0.172	0.340
3	0.883	0.190	0.279	0.219
4	1.589	0.158	0.632	0.202
5			1.002	0.141

Table 10. Estimated regression coefficients in the proportional odds model for the latent class probabilities

Covariate	*Results for quartiles*				*Results for fixed cut-offs*			
	Estimate	*Standard error*	*t-statistic*	*p-value*	*Estimate*	*Standard error*	*t-statistic*	*p-value*
Intercept 1	3.506	0.444	7.896	0.000	5.574	0.752	7.417	0.000
Intercept 2	1.125	0.306	3.677	0.000	2.479	0.434	5.709	0.000
Intercept 3	−0.260	0.375	−0.693	0.488	1.021	0.356	2.869	0.004
Intercept 4					−0.452	0.312	−1.451	0.147
Age 1–3 years	−3.974	1.030	−3.859	0.000	−4.372	0.803	−5.446	0.000
Age 4–6 years	−2.905	0.553	−5.252	0.000	−2.231	0.470	−4.749	0.000
Age 7–10 years	−0.801	0.424	−1.892	0.058	−0.908	0.407	−2.231	0.026
Age 11–20 years	−0.734	0.318	−2.306	0.021	−0.597	0.316	−1.889	0.059
Age 21–40 years	−0.315	0.290	−1.089	0.276	−0.337	0.288	−1.168	0.243
UK, D, NL	−0.553	0.256	−2.156	0.031	−0.582	0.249	−2.340	0.019
Other countries	−3.650	0.395	−9.238	0.000	−3.326	0.345	−9.646	0.000

country of the publisher (the reference category is published in the USA), in the proportional odds model (5) for the latent class probabilities. The estimated coefficients suggest that the probability of being in a higher class (i.e. having a higher impact) increases with a journal's age and, compared with journals published in the USA, it is slightly lower for those published in the UK, Germany and the Netherlands, and is much lower for those published in all other countries.

Table 11 presents estimates of the discriminant indices α_j in equation (2), which allow us to assess the quality of each indicator as a measure of the impact of a journal. Table 11 also presents the estimates of the difficulty parameters β_{jv}. It is worth noting that we reach the same conclusion under both types of discretization: the indicator that best measures the impact of a journal is IF5 because it corresponds to the highest estimate of the discrimination parameter. Moreover, the estimates of the discrimination parameter for the other indicators are quite similar under both types of discretization. This agrees with the result in Chang *et al*. (2010), based on the ISI–Thomson Reuters database of citations from all fields in the sciences and social sciences, that AIS does not add much compared with more traditional indicators such as IF5.

We then assign each journal to a latent class by using the rules that were discussed in Section 2.3. Under the first type of discretization (quartiles), we assign 159 journals to the first class (lowest impact), 138 to the second, 79 to the third and 69 to the fourth (highest impact). The

Table 11. Estimated discriminant indices and difficulty parameters

j	*Results for quartiles*				*Results for fixed cut-offs*				
	$\hat{\alpha}_j$	$\hat{\beta}_{j1}$	$\hat{\beta}_{j2}$	$\hat{\beta}_{j3}$	$\hat{\alpha}_j$	$\hat{\beta}_{j1}$	$\hat{\beta}_{j2}$	$\hat{\beta}_{j3}$	$\hat{\beta}_{j4}$
1	3.124	0.000	2.526	4.670	6.862	0.000	2.507	3.754	6.095
2	5.132	0.709	3.195	4.924	13.765	1.528	2.825	4.842	6.442
3	3.234	0.614	2.950	4.921	6.594	1.713	4.278	5.443	6.915
4	3.523	−2.818	−0.276	3.200	6.743	−4.026	−0.705	2.662	4.750

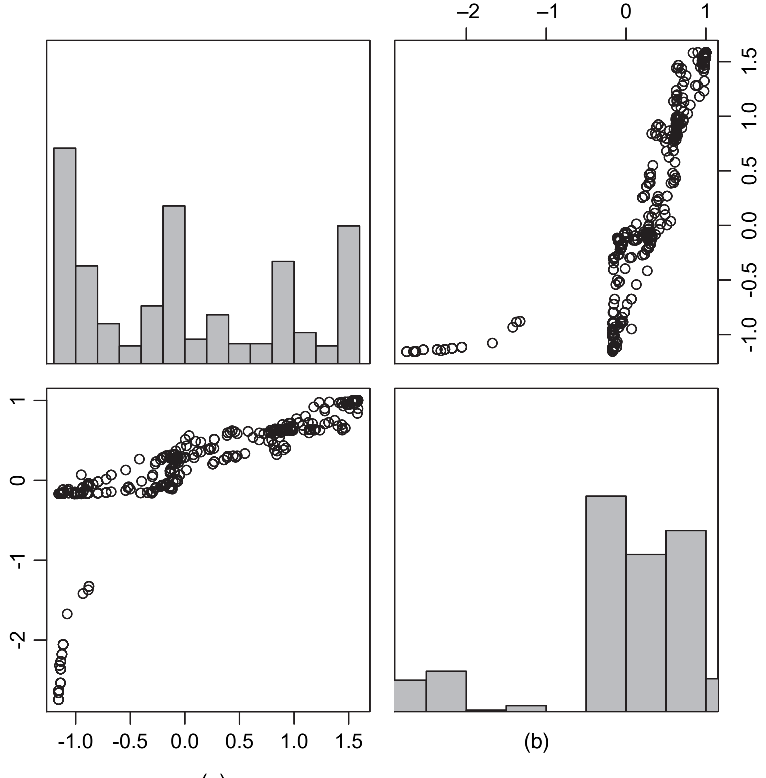

Fig. 2. Scatter plot of the predicted latent variables under the discretization based on (a) quartiles and (b) fixed cut-offs

journals that were assigned to the top two classes are listed in Tables 1 and 2, together with information of interest, such as the estimated posterior probabilities and the predicted impact under both models. Fig. 2 presents the scatter plot of the predicted impact of all journals. These results show that, although we are assuming a discrete distribution for the latent variable with a relatively small number of support points, there is a high variety of different predicted values, confirming our previous remarks in Section 2.3. This is because for some journals the estimated posterior probabilities are not strongly unbalanced towards a certain class, as confirmed by the plots in Fig. 3.

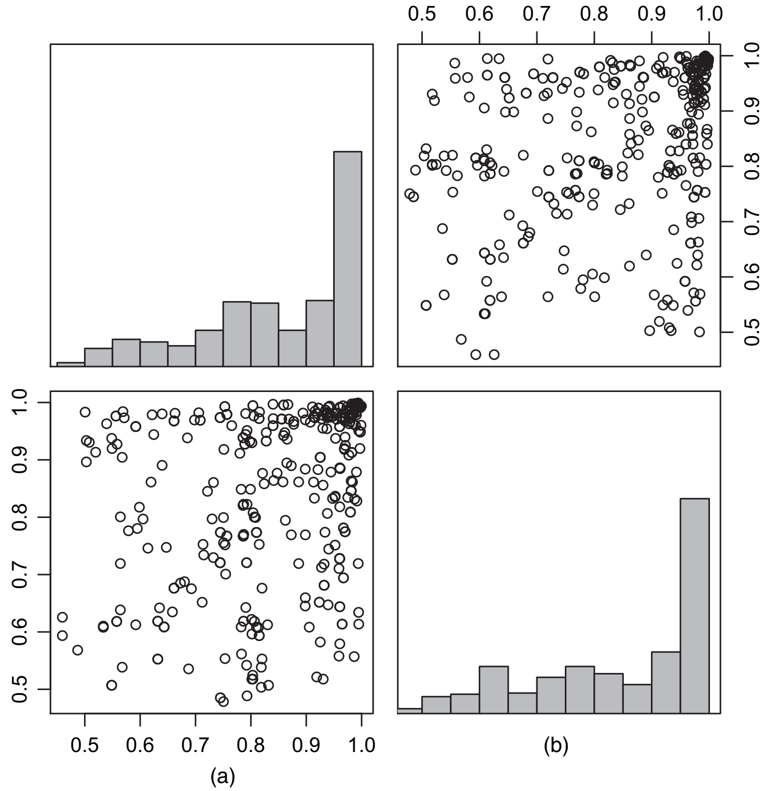

Fig. 3. Scatter plot of the maximum posterior probabilities under the discretization based on (a) quartiles and (b) fixed cut-offs

On the basis of the above journal clustering, Table 12 reports the average conditional correlation between the indicators given the latent class under the two types of discretization, together with the observed (marginal) correlation between the indicators. Note that the conditional correlation between the pairs of indicators for which we assume local independence is very low compared with both the observed correlation and the conditional correlation between the pair of indicators (IF and IF5) for which we admit local dependence. This provides indirect support for the assumptions underlying the models adopted.

5. Conclusions

We propose an approach to ranking scientific journals in a given list based on a latent variable model for polytomous item response data. The latent variables, which are assumed to be discrete and are interpreted as the unobserved scholarly impact of the journals in the list, are predicted on the basis of a set of bibliometric indicators that are discretized to avoid strong parametric assumptions.

Our approach has the advantage of relying on a well-defined statistical model, which simplifies the interpretation of the results and inference. The model is semiparametric in nature as it does not need parametric assumptions on the distribution of the latent variables. This is instead approximated by a discrete distribution with a finite number of support points corresponding to a set of latent classes, i.e. groups of journals with similar unobservable characteristics.

Table 12. Correlation matrix of the observed indicators and average of the conditional correlation given the assigned cluster

Indicator	*Results for the following indicators:*			
	IF	*IF5*	*AIS*	*h-index*
Observed				
IF	1.000	0.899	0.692	0.556
IF5	0.899	1.000	0.795	0.579
AIS	0.692	0.795	1.000	0.485
h-index	0.556	0.579	0.485	1.000
Conditional (quartiles)				
IF	1.000	0.724	0.147	0.108
IF5	0.724	1.000	0.265	0.157
AIS	0.147	0.265	1.000	0.004
h-index	0.108	0.157	0.004	1.000
Conditional (fixed cut-offs)				
IF	1.000	0.671	0.120	0.010
IF5	0.671	1.000	0.256	0.094
AIS	0.120	0.256	1.000	0.045
h-index	0.010	0.094	0.045	1.000

Journals are assigned to a latent class on the basis of a maximum *a posteriori* probability rule. The posterior mean of the latent variable provides a prediction on a single continuous scale of the impact of each journal, so journals can be ranked, the distance between any pair of them can be compared and journals can be clustered into any arbitrary number of classes of a given size. One can also assess the discriminant power of each indicator, i.e. the sensitivity and reliability of each indicator in its relationship to the latent variable. For example, in our data we find that IF5 is more reliable than IF, AIC and the h-index as an indicator of the impact of a journal. Finally, our approach has the advantage of easily handling missing values in the available indicators.

A key aspect of our approach is that it requires a preliminary discretization of the available bibliometric indicators. It is therefore important to assess the sensitivity of the results of an analysis to the discretization adopted. As shown in our application, this can be assessed by replicating the analysis with different discretizations and then comparing the results that are obtained. In our empirical application, the results appear to be fairly robust to the choice of discretization.

Our analysis can be easily extended to handle the case where different indicators have different numbers of categories (to include, for instance, binary indicators), and to employ discriminant indices which are category dependent. Finally, our method could be applied to other lists of journals in different fields.

Acknowledgements

We thank Sergio Benedetto, Tullio Jappelli, Daniele Terlizzese, the Associate Editor, two referees and seminar participants at the Einaudi Institute for Economics and Finance 2013 and the 5th Italian Congress of Econometrics and Empirical Economics for helpful comments.

References

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

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proc. 2nd Int. Symp. Information Theory* (eds B. N. Petrov and F. Csàki), pp. 267–281. Budapest. Akademiai Kiàdo.

Bacci, S., Bartolucci, F. and Gnaldi, M. (2014) A class of multidimensional latent class irt models for ordinal polytomous item responses. *Communs Statist. Theor. Meth.*, **43**, 787–800.

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

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A. and Peracchi, F. (2014) Bibliometric evaluation vs. informed peer review: evidence from Italy. *Res. Poly*, to be published.

Biernacki, C., Celeux, G. and Govaert, G. (1999) An improvement of the nec criterion for assessing the number of clusters in a mixture model. *Pattn Recogn. Lett.*, **20**, 267–272.

Bollen, J., van de Sompel, H., Hagberg, A. and Chute, R. (2009) A principal component analysis of 39 scientific impact measures. *PLOS ONE*, **4**, article 6.

Celeux, G. and Soromenho, G. (1996) An entropy criterion for assessing the number of clusters in a mixture model. *J. Classificn*, **13**, 195–212.

Chang, C.-L., Maasoumi, E. and McAleer, M. (2012) Robust ranking of journal quality: an application to economics. *Working Paper 813*. Institute of Economic Research, Kyoto University, Kyoto.

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

Colombi, R. and Forcina, A. (2001) Marginal regression models for the analysis of positive association of ordinal response variables. *Biometrika*, **88**, 1007–1019.

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

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

Douglas, R., Fienberg, S. E., Lee, M. T., Sampson, A. R. and Whitaker, L. R. (1991) Positive dependence concepts for ordinal contingency tables. In *Topics in Statistical Dependence* (eds H. W. Block, A. R. Sampson and T. H. Savits), pp. 189–202. Hayward: Institute of Mathematical Statistics.

Fraley, C. and Raftery, A. E. (2002) Model-based clustering, discriminant analysis, and density estimation. *J. Am. Statist. Ass.*, **97**, 611–631.

Garfield, E. (2006) The history and meaning of the journal impact factor. *J. Am. Med. Ass.*, **295**, 90–93.

Goodman, L. A. (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

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

Heckman, J. and Singer, B. (1985) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Hirsch, J. (2005) An index to quantify an individual's scientific research output. *Proc. Natn. Acad. Sci. USA*, **102**, 16569–16572.

Lazarsfeld, P. F. and Henry N. W. (1968) *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lindsay, B., Clogg, C. and Greo, J. (1991) Semiparametric estimation in the rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Am. Statist. Ass.*, **86**, 96–107.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*, 2nd edn. New York: Wiley.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

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

Palacios-Huerta, I. and Volij, O. (2004) The measurement of intellectual influence. *Econometrica*, **72**, 963–977.

Plackett, R. (1965) A class of bivariate distributions. *J. Am. Statist. Ass.*, **60**, 516–522.

Samejima, F. (1969) Estimation of ability using a response pattern of graded scores. *Psychmetr. Monogr.*, **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. *Ann. Statist.*, **6**, 461–464.

Zimmermann, C. (2012) Academic Rankings with RePEc. *Working Paper 2012-023A*. Federal Reserve Bank of St Louis, St Louis.