Uncovering Excellence in Academic Rankings
A case study on the Shanghai Ranking

Catherine Dehon † Alice McCathie ‡ Vincenzo Verardi §

May 28, 2009

Abstract

In this paper we examine whether the quality of academic research can be accurately captured by a single aggregated measure such as a ranking. With Shanghai University’s Academic Ranking of World Universities as the basis for our study, we use robust principal component analysis to uncover the underlying factors measured by this ranking. Based on a sample containing the top 150 ranked universities, we find evidence that, for the majority of these institutions, the Shanghai rankings reflect not one but in fact two different and uncorrelated aspects of academic research: overall research output and top-notch researchers. Consequently, the relative weight placed upon these two factors determines to a large extent the final ranking.

* We would like to thank Nadine Rons, Ellen Hazelkorn and Marjorie Gassner for their insightful comments. All remaining errors are the authors' own.

† Université Libre de Bruxelles, ECARES (European Centre for Advanced Research in Economics and Statistics) and CKE (Centre for Knowledge Economics). Email: cdehon@ulb.ac.be

‡ Université Libre de Bruxelles, ECARES (European Centre for Advanced Research in Economics and Statistics) and CKE (Centre for Knowledge Economics). Email: amccathi@ulb.ac.be

§ University of Namur, CRED (Centre for Research in the Economics of Development). Université Libre de Bruxelles, ECARES (European Centre for Advanced Research in Economics and Statistics) and CKE (Centre for Knowledge Economics). Email: vverardi@ulb.ac.be. Vincenzo Verardi is associate researcher with FNRS and gratefully acknowledges their financial support.
1 Introduction

The ranking of higher education institutions (HEIs) has garnered a lot of attention in recent times. Prospective students, especially postgraduates, use rankings to get an idea of a university’s relative performance. This in turn encourages public bodies to pay attention to rankings when allocating funds to HEIs. With increased competition between institutions due to higher participation rates and greater student mobility (Marcotte et al, 2007), university rankings appeal to students and governments alike because they provide information that would otherwise be difficult and costly to acquire. This widespread influence is putting pressure on schools to tailor themselves to better meet the criteria relevant to rankings in order to attract both students and funds. However, despite growing popularity, the ranking of universities remains a controversial issue and has been widely debated. Dehon et al (2009) offer a global view of these issues. Indeed, the very idea of summing up the various characteristics of an institution in a single measure of performance is questionable. And as rankings proceed from the aggregation of information, their results can vary due to the chosen methodology and criteria. Given the scope of their influence and their various drawbacks it is important to understand how these rankings are designed and what exactly they are measuring.

This is precisely our aim here: to determine whether academic excellence, or more specifically the quality of research, can be measured by a single factor reflected in a ranking. With this objective in mind we choose to take a closer look at the Academic Ranking of World Universities (ARWU). First published in 2003, the ARWU was created by a group of researchers at the University of Shanghai, and is now one of the best known international ranking of universities.\(^1\) While its initial purpose was to ascertain the relative position of Chinese universities internationally, the ARWU has since attracted a lot interest from around the world. It now receives considerable press attention when it is published once a year. One of the reasons behind this relative popularity, as well as our decision to choose the ARWU as the basis of this study, is the transparency and simplicity of the methodology and their use of objective criteria. Their website features the scores used to compute the rankings, however the raw data is not available. The missing data means that the results cannot be reproduced, which is unfortunate, as Razvan (2007) notes, since reproducibility is a principal requirement for any scientific method.

In order to uncover the underlying factor or factors measured by the “Shanghai rankings”, we perform a principal component analysis (PCA) on the 2008 ARWU results. This allows us to identify a single significant underlying factor to the data, a result which seems to indicate that the ARWU’s various criteria each measure a similar aspect of research excellence. However, we observe at this stage the presence of an extreme value in the data, Harvard University. We furthermore suspect that the sample contains other outliers, thereby invalidating the results of the PCA. To address this issue we need a robust methodology. The Reweighted Minimum Covariance Determinant (RMCD) es-

\(^1\)See http://ed.sjtu.edu.cn/ranking.htm
imator is a robust estimator of the sample covariance matrix, which we use to compute the underlying factors. The results of this robust PCA suggest that, for the vast majority of the universities in the sample, the ARWU does not measure a single, common factor of excellence, but actually combines two different and uncorrelated aspects of research performance: overall output on one hand, top-notch researchers on the other.

The outline of the paper is as follows. We explain in section 1 the methodology used by Shanghai University to construct its rankings. In section 2 we present the results of the principal component analyses, both from the regular approach and the robust method. Based on these results, we propose in section 3 alternative rankings by varying the weights given to the component-variables identified by the robust PCA. Finally, we present our concluding remarks in section 4.

2 The Shanghai Rankings: Methodology and Results

The Shanghai rankings draw on six different criteria to measure academic performance. In each category the best performing university is given a score of 100 and becomes the benchmark by which the scores of all the other universities are measured. Universities are then ranked according to the overall score they obtain, which is simply a weighted average of their individual category scores. The variables “Alumni” and “Award” measure the number of Nobel prizes and Fields medals won by a university’s alumni (“Alumni”) or current faculty members (“Award”). The next three variables, “HiCi”, “N&S” and “PUB”, are said to reflect the quality of research. “HiCi” is the number of highly cited researchers, “N&S” is the number of articles published in the journals Nature and Science, and “PUB” is the number of articles indexed in the Science Citation Index-expanded and the Social Science Citation Index. The sixth and final variable, “PCP”, is a weighted average of the scores obtained in the previous five categories, divided by the number of current full-time equivalent academic staff members. The variables “Award”, “HiCi”, “N&S” and “PUB” each make up 20% of final score, while “Alumni” and “PCP” are each given the slightly lower weight of 10%.

Despite its popularity, the Shanghai rankings have come under some criticism regarding both their methodology and choice of variables (Liu and Cheng, 2005 and van Raan, 2005b). Vincke (2009) notes that using an averaged score to measure performance has a determining influence on the ranking. With respect to the choice of variables, Shanghai University uses only a limited set of criteria, which measure academic performance solely in terms of research excellence, to rank a wide range of universities. This “one size fits all approach” fails to capture the specific characteristics of a university and ignores the objectives an institution pursues outside of research, such as education and a social mission. In terms of its criteria, the ranking is biased in favor of science and
technology and almost totally disregards other fields such as the arts and humanities. Thus, schools with strong scientific departments fare much better in the rankings than schools that specialize in the arts, humanities or social sciences. The ARWU also favors English-speaking universities as English is the predominant language of academic publications. Van Raan (2005a) points out these biases and warns against the misuse of overly simple bibliometric indicators. Finally, the ARWU does not take into account the effect of size on performance. Zitt (2007) notes that ninety percent of criteria used in rankings are size-dependant. Indeed, the Shanghai rankings essentially measure overall production and not efficiency, an approach that favors large universities. And while they do include one variable to this effect (“PCP”), it is rendered almost useless as it is only computed for the universities which survive pre-selection based on their performance with respect to the other criteria.

3 Uncovering the underlying factors of influence

3.1 Principal Component Analysis

In order to determine the underlying factors measured by the Shanghai rankings, we perform a principal component analysis (PCA). PCA allows you to reduce the dimension of analysis by producing a set of orthogonal factors called principal components, which underlie the data and capture the correlations that exist between the variables. These principal components are linear combinations of the standardized variables and are obtained by computing the eigenvectors of the sample covariance or correlation matrix. In standardized PCA, the eigenvectors are obtained from the correlation matrix.

We choose here to restrict our analysis to the universities ranked in the top 150. This decision stems from the changing relationship between overall score and rank. Figure 1 clearly shows that while a downward slopping convex curve can be fitted for the first 150 universities, the relationship beyond this point becomes linear and is characterized by a very shallow slope. Indeed, for the universities ranked beyond the 150 mark, the drop in score which prompts a drop in rank is quite small. Thus, as we go down the ranking, the differences between universities become less meaningful. Shanghai University’s choice of normalization scheme, by computing each university’s intermediate scores relative to those of the highest scoring university, accentuates this. We furthermore drop the variable “PCP” from our analysis as it makes for an awkward combination with the other variables which are measured in absolute terms. “PCP” also introduces a selection bias in the sample as it is only computed for the universities that survive pre-selection on the basis of the five other criteria, which are themselves influenced by the size.

The principal components obtained from this reduced data set are featured in Table 1. The variance captured by each principal component is given by its associated

\footnote{In standardized PCA, the eigenvectors are obtained from the correlation matrix.}

\footnote{Computed after removing the variable “PCP” and transferring its weight equally between the remaining five variables.}
eigenvalue. As a general decision rule a factor is kept if it explains a higher than average proportion of the overall variance, or in other words, if its associated eigenvalue is greater than one.\footnote{Since we are working here with standardized variables, the overall variance here is equal to five, the number of variables.} From Table 1 we see that only the first factor, $\phi_1$, fits this decision rule, with an associated eigenvalue of 3.4. This component captures 64% of the variance in our data set and is given by:

$$\phi_1 = 0.42 \times Alumni + 0.44 \times Award + 0.48 \times HiCi + 0.50 \times N&S + 0.38 \times PUB.$$  

<table>
<thead>
<tr>
<th>Component $\phi_i$</th>
<th>Eigenvalue $\lambda_i$</th>
<th>Cumulative explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>3.40</td>
<td>68%</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>0.87</td>
<td>86%</td>
</tr>
<tr>
<td>$\phi_3$</td>
<td>0.41</td>
<td>94%</td>
</tr>
<tr>
<td>$\phi_4$</td>
<td>0.19</td>
<td>98%</td>
</tr>
<tr>
<td>$\phi_5$</td>
<td>0.12</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 1: Principal component analysis: results

The correlation matrix of Table 2 shows that the first principal component is strongly positively correlated with each of the initial variables. $\phi_1$ thus represents a common underlying factor to all five variables. This indicates that each variable is measuring the...
same common factor of performance. As a result, any university that scores highly in one category is expected to also perform well in the others. These findings suggest that the ARWU is indeed based on a set of variables that accurately reflect a common aspect of research performance and are thus successful in measuring excellence.

<table>
<thead>
<tr>
<th>$\phi_1$</th>
<th>Alumni Award</th>
<th>HiCi</th>
<th>N&amp;S</th>
<th>PUB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.78</td>
<td>0.81</td>
<td>0.89</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table 2: Correlation between the first principal component and the initial set of variables

However, if we consider the influence of each university on the total inertia of the first factor, the above results are cast in a different light. Contribution of a university $i$ to a factor $\phi_1$ is measured by: $CTR_i = \frac{\phi_1^2}{n} \times \frac{1}{\lambda_1}$. In this first PCA, the top ten universities together contribute over 60% of the first component, with Harvard alone contributing 18%.\(^5\) By scoring well above all the other universities in the sample, Harvard is an important source of variance. And since the first principal axis\(^6\) is fitted to the data cloud so as to capture maximum variance, it makes sense to observe that Harvard overly influences the first component. The presence of such an extreme value in the data set invalidates the results of this analysis. Our next step is thus to identify the other outliers that might be present in the sample.

### 3.2 Detection of outliers

The results from the principal component analysis revealed the presence of an extreme value in the data set, Harvard University. This leads us to suspect that the sample may be contaminated by outliers. In order to identify these outliers, we use a well-known measure of outlyingness, the Mahalanobis distance (Mahalanobis, 1936), which is defined as follows:

$$d_i = \sqrt{(X_i - \mu)\Sigma^{-1}(X_i - \mu)'},$$

where $X_i$ is a $(1 \times p)$ vector containing the observations for individual $i$, $\mu$ is the $(1 \times p)$ multivariate location vector and $\Sigma$ is the $(p \times p)$ covariance matrix. The Mahalanobis distance measures how far an observation lies from the centre of gravity of the multivariate data cloud. Observations can then be ranked according to this distance and classed as outliers beyond a well-specified cut-off point. To compute the Mahalanobis distances of our various data points, we need estimates of both the mean vector $\mu$ and covariance matrix $\Sigma$. Using the empirical mean and empirical covariance matrix is unsatisfactory as both these measures are sensitive to outliers. This could lead to certain outliers going

\(^5\)Obviously, if each university were to contribute an equal share in a factor, that contribution would be 0.67%.

\(^6\)The first principal component is measured on the first principal axis
undetected due to the influence of other even more extreme values. This is known as the masking effect. To avoid this we need robust estimates of both $\mu$ and $\Sigma$. The Minimum Covariance Determinant (MCD) estimator provides us with robust estimates of both parameters (Rousseeuw, 1985). The MCD estimator is simply the covariance matrix of the subsample with the smallest generalized variance that contains 50% of the observations. Defined as the determinant of the covariance matrix, generalized variance is a one-dimensional assessment of the multivariate spread. Thus finding the MCD estimator simply entails minimizing the determinant of the covariance matrix across possible subsamples. Using the robust Mahalanobis distance measure and the 95th percentile of a chi squared distribution\(^7\), we identify seven-thirty outliers from the 150 universities in the sample.\(^8\) Figure 2 shows the Mahalanobis distances for all our data points. As expected, Harvard University is clearly identified as an outlier. In fact, 29 of the thirty top ranked universities are outliers. It is also interesting to note that 36 of the 37 universities classed as outliers are ranked in the top 100.

![Figure 2: Robust Mahalanobis distances](image)

### 3.3 Robust Principal Component Analysis

Having identified the outliers present in the data set, we move forward with a robust principal component analysis. One option is to work directly with the MCD estimator obtained previously and compute its eigenvalues and associated eigenvectors to deter-

\(^7\)Under the normality assumption, the Mahalanobis distances are distributed as the square root of a $\chi^2$ distribution with five degrees of freedom.

\(^8\)The STATA code is available from the authors upon request (Verardi and Dehon, 2009)
mine the principal components. Alternatively, we can use the Reweighted Minimum Covariance Determinant (RMCD) estimator, which is a robust covariance matrix computed on our sample after downweighting the outlying observations. We choose the second approach, which yields results that are both more efficient and more stable. Table 3 reports the results of the robust principal component analysis based on the RMCD estimator. Using the same decision rule as before, we see that the first two factors appear significant. They are given by:

\[ \phi_1 = -0.03 \times Alumni + 0.02 \times Award + 0.62 \times HiCi + 0.60 \times N&S + 0.50 \times PUB \]

\[ \phi_2 = 0.69 \times Alumni + 0.71 \times Award - 0.04 \times HiCi + 0.14 \times N&S - 0.11 \times PUB. \]

<table>
<thead>
<tr>
<th>Component ( \phi_i )</th>
<th>Eigenvalue ( \lambda_i )</th>
<th>Cumulative explained variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \phi_1 )</td>
<td>1.97</td>
<td>39%</td>
</tr>
<tr>
<td>( \phi_2 )</td>
<td>1.45</td>
<td>68%</td>
</tr>
<tr>
<td>( \phi_3 )</td>
<td>0.70</td>
<td>82%</td>
</tr>
<tr>
<td>( \phi_4 )</td>
<td>0.54</td>
<td>93%</td>
</tr>
<tr>
<td>( \phi_5 )</td>
<td>0.34</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3: Results of the robust principal component analysis

Together they explain around 68% of the overall variance. Furthermore, as we see in Table 4, each of these two factors is correlated with a different subset of the original variables. The first component, \( \phi_1 \), is strongly positively correlated with “HiCi”, “N&S” and “PUB”, variables that essentially measure the volume of publication and the number of researchers, and are thus highly influenced by the size of a university. The second component, \( \phi_2 \), is strongly positively correlated with “Award” and “Alumni”. These variables account for the number of top-notch researchers and as such reflect the quality of research conducted at the highest level. The Shanghai rankings thus proceed from the aggregation of two uncorrelated factors of research excellence: overall productivity on one hand and top-notch researchers on the other. By combining orthogonal measures of performance in a single overall score, the ARWU is in effect drowning out the information that each factor contains.

4 Alternative Rankings with Varying Weights

Our findings suggest that the Shanghai rankings are sensitive to the relative weight it attributes to each of the variables. To verify this we compute alternative rankings with different weighting schemes. We do this for two different groups of universities: the top
Table 4: Robust PCA: Correlation between the first two principal components and the initial set of variables

<table>
<thead>
<tr>
<th></th>
<th>Alumni</th>
<th>Award</th>
<th>HiCi</th>
<th>N&amp;S</th>
<th>PUB</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>-0.04</td>
<td>0.03</td>
<td>0.87</td>
<td>0.85</td>
<td>0.70</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>0.82</td>
<td>0.85</td>
<td>-0.05</td>
<td>0.16</td>
<td>-0.13</td>
</tr>
</tbody>
</table>

10 and a selection of European universities. Having uncovered two different underlying factors to the set of variables, each correlated with a different subset of the original variables, we group these accordingly to compute the new scores, and consider the following weighting scheme:

$$SCORE_i = w_i \times (\text{Alumni} + \text{Award}) + (1 - w_i) \times (\text{HiCi} + \text{N&S} + \text{PUB}),$$

with $w_i = 0, 0.1, \ldots, 1$.

The alternative rankings obtained are plotted in Figures 3 and 4. Firstly, we notice from Figure 3 that changing the variable weights has almost no effect on the top ten ranked universities. Indeed, the rankings remain relatively stable across the different weighting schemes. This indicates that for the top ten ranked universities, all five variables truly measure the same thing and there is thus a single underlying factor.

This however does not hold true for the majority of the universities in our sample. If we look at the alternative rankings obtained for several European universities in Figure 4, we see that different weighting schemes yield very different results. The Vrij Universiteit Amsterdam for example, is the only university whose ranking benefits from putting a smaller weight on the “Nobel” variables. This is simply because the VU Amsterdam has never won a Nobel Prize or Fields medal. Of the six universities plotted here, the University of Moscow (ARWU rank 71) and the Ecole Normale Supérieure of Paris (ARWU rank 74) have the highest rank in the ARWU. Coincidentally, their rankings in this experiment are those most sensitive to the differing weighting schemes. By shifting the entire weight from the “volume of publication” variables to the “Nobel” variables, we can make them each drop over 200 places and fall to the bottom of the table. The university whose ranking is the least affected in our small sample of European institutions is the University of Geneva. The variability in the rankings generated here show the extent to which the chosen methodology and criteria have a determining effect on the overall result.
5 Concluding Remarks

The academic world and its different stakeholders are becoming ever more concerned with the assessment of higher education. Rankings are sprouting up in response to this
demand for ways of comparing the different merits of institutions. Since rankings proceed from the aggregation of various and somewhat diverse characteristics, they are subject to much criticism. In this paper we examine whether the quality of academic research can be accurately captured by a single aggregated measure such as a ranking. Our results highlight some of the conceptual and methodological problems with ranking higher education institutions, and warn against the use of one-dimensional measures which can be misleading due to oversimplification. With Shanghai University’s Academic Ranking of World Universities as the basis for our study, we use robust principal component analysis (PCA) to uncover the underlying factors measured by this ranking. While PCA is a relatively straightforward and commonly-used method of dimension reduction, it does have certain limitations. In particular, it is a purely descriptive method for quantitative data only, that solely considers the linear dependence between variables. Our analysis shows that for the majority of the institutions in our sample, their ranking in the ARWU proceeds from the combination of two different and uncorrelated aspects of research performance. The first factor assesses research conducted at the highest level, as measured by alumni and faculty recipients of a Nobel Prize or Fields medal, while the second factor measures research in terms of output. This aggregation of uncorrelated factors of performance leads to the loss of some of the information they contain. As a result, changing the relative weight placed upon these two factors significantly alters the ranking.

There is a well-known saying that tells us “if you can’t beat ’em, join ’em”. So if rankings are to be a permanent part of the academic landscape, we should encourage the ongoing debate and establish further guidelines and criteria on how best to conduct and interpret higher education quality assessments.

6 Bibliography

Razvan, F. (2007), Irreproducibility of the results of the Shanghai academic ranking of world universities, Scientometrics, 72, No. 1: 25-32.


Verardi, V. and Dehon C. (2009), Multivariate Outlier Detection in Stata, working paper.
