Changes in Occupational Skills - A Case Study Using Non-negative Matrix Factorization

Wei Lee Woon(✉), Zeyar Aung, Wala AlKhader, Davor Svetinovic, and Mohammad Atif Omar

Institute Center for Smart and Sustainable Systems,
Masdar Institute of Science and Technology, P.O. Box 54224, Abu Dhabi, UAE
{wwoon,zaung,wabedalkhader,dsvetinovic,momar}@masdar.ac.ae

Abstract. Changes in the skill requirements of occupations can alter the balance in the numbers of high, middle and low-skilled jobs on the market. This can result in structural unemployment, stagnating income and other unforeseen social and economic side effects. In this paper, we demonstrate the use of a recent matrix factorization technique for extracting the underlying skill categories from O*NET, a publicly available database on occupational skill requirements. This study builds upon earlier work which also focused on this database, and which indicated that changes in skill requirements were in response to increased automation which unevenly affected different segments of the job market. In this paper we refine the methodological underpinnings of the earlier work and report some preliminary results which already show great promise.

Keywords: Non-negative matrix factorization · Data mining · Source separation · Empirical research · Job characteristics

1 Introduction

1.1 Background and Related Work

Technological advances have long had a disruptive effect on the job market. While the long term effects have generally been increased productivity and efficiency, this has often been at the expense of significant, if transient, social imbalance.

However, the pace of recent advances in digital technologies and automation have been unprecedented. IBM's "Watson", which competed with and defeated the best human competitors of the *Jeopardy* quiz show, is just one example of a range of technologies that are precipitating a broad shift towards increased automation in both a greater number and variety of jobs. Furthermore, it apparent that these changes will continue or even accelerate in the foreseeable future. The socio-economic impacts of these trends are far reaching - middle and low skill jobs are disappearing while labor force participation and median incomes have fallen [1]. It is unclear how these changes will develop in the long run but there is a pressing need to study the underlying factors so that informed mitigation strategies can be formulated.

S. Arik et al. (Eds.): ICONIP 2015, Part III, LNCS 9491, pp. 627–634, 2015.
DOI: 10.1007/978-3-319-26555-1_71

An earlier paper [2] (henceforth referred to as MC) examined how the skill content of jobs had evolved in the period between 2006 and 2014 using O*NET, a publicly available and comprehensive occupational skill requirements database[1] compiled by the US government. The findings of that study support the notion that substitution effects will remove some skills from occupations, complementarity effects will amplify other skills, and skills that are orthogonal will be amplified due to Baumols Cost Disease [3].

1.2 Motivations and Objectives

Key to successfully understanding these issues is the ability to detect and study the underlying skill "dimensions", i.e. groups of skill or ability elements which co-occur repeatedly across multiple occupations.

In [4], these were manually constructed based on domain knowledge, while in MC factor analysis (FA) was used to affect data driven skills aggregation. While the results of these studies were already very insightful, manual extraction is a highly subjective process while FA is based on a number of assumptions, in particular that the underlying factors are Gaussian distributed and zero mean. These are difficult to substantiate at best; in fact, the ratings provided in O*NET range from 1 to 5, while a summary inspection of the importance levels reveal a mix of different distributions, many of which are clearly not Gaussian (some examples are shown in Fig. 1). An additional issue is that the factor loadings include negative coefficients which are very difficult to interpret.

This paper addresses these concerns by, on the one hand, retaining the empirical approach to factor elicitation while on the other hand identifying and testing suitable alternatives to factor analysis. To evaluate the usefulness of the proposed method, we repeat elements of the analysis performed in MC and report on the findings and observations.

2 Methods and Data

2.1 O*NET

The Occupational Information Network (O*NET) is a publicly available database of occupations developed for the US Department of Labor and is the successor to the better-known Dictionary of Occupational Titles. O*NET contains detailed information on more than 950 US occupations, including the tasks associated with the various occupations, required knowledge, skills and abilities, typical work activities and the contexts in which the occupations are commonly performed. To facilitate comparison with earlier work, we utilize data matrices constructed using the combined importance levels for three main job characteristics: *Abilities*, *Skills* and *Work Activities*.

Also, note that O*NET only documents occupation types and not the demand for or frequency of these occupations. While this is a limitation, it also imposes

[1] http://www.onetonline.org.

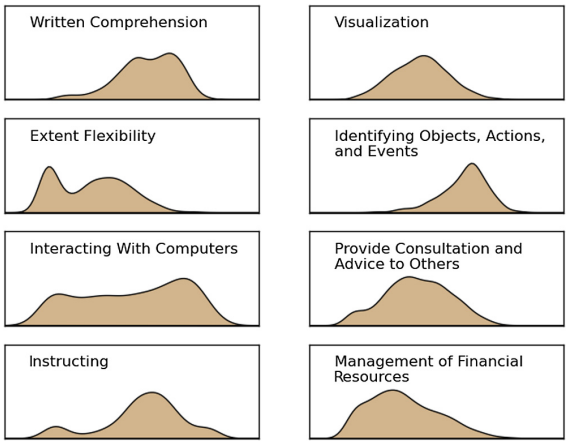

Fig. 1. Probability densities for a selection of skills/abilities/activities (2006 levels)

an emphasis on the study of *intensive* changes in occupations, *viz* how the actual composition of jobs is changing. This is interesting because most existing studies have focused on *extensive* changes even though it is known that both types of changes occur and are strongly affected by technological progress [5].

However, the data mentioned above is of extremely high dimensionality (even using only a subset of the job characteristics has resulted in a 128-dimensional data set). This is why it is particularly important to identify the most informative combinations of these dimensions if this data is to be effectively analyzed.

2.2 Factor Analysis

FA was used as a refinement to the approach taken in [4], where skills dimensions were empirically identified based on the statistical properties of the importance levels. FA models data as a linear combination of Gaussian distributed, uncorrelated *factors*. It is similar to PCA but with the addition of independent, Gaussian distributed error terms to each input variable. For a set of observed variables $x_i, \ldots, x_p$ with expected values $\mu_i, \ldots, \mu_p$:

$$x_i - \mu_i = l_{i1}F_1 + l_{i2}F_2 + \ldots + l_{ik}F_k + \epsilon_i, \tag{1}$$

where l_{ij} is the loading of the jth factor on the ith variable and F_j is factor j. To improve the readability of the resulting factor loadings, varimax rotation was subsequently applied to the loadings matrices. For conciseness, we will subsequently refer to this combination of factor analysis with varimax as FA.

2.3 Non-negative Matrix Factorization

As mentioned previously, the main objective of this work was to extend the methodological basis of MC by identifying and testing suitable alternatives

to FA. The motivations for this are (i) The variables of interest (skill importance levels) are non Gaussian distributed (ii) Negative loadings produced by FA can be difficult to interpret.

One method which addresses both these concerns is Non-Negative Matrix Factorization or NMF [6]. NMF is similar to FA it that it aims to express a matrix of data as a linear combination of a set of basis vectors and a transformed data set:

$$V \approx WH, \tag{2}$$

where V is the matrix of row vectors containing the input data, W is the transformed data set and H is the basis set which defines a linear combination of the columns of W (to permit easier discussion we henceforth refer to W and H as the *factor* and *loading* matrices respectively). However, unlike FA, NMF makes no assumptions about Gaussianity or even orthogonality of the underlying factors. Instead, the only constraint is that all three matrices V, W and H are non-negative.

Since (2) does not have a unique or closed form solution, it is typically solved iteratively using the following multiplicative update terms [6]:

$$H_{bj}^{k+1} = H_{bj}^{k} . \frac{((W^k)^T V)_{bj}}{((W^k)^T W^k H^k)_{bj}} \tag{3}$$

$$W_{ia}^{k+1} = W_{ia}^{k} . \frac{(V(H^{k+1})^T)_{ia}}{(W^k H^{k+1}(H^{k+1}))_{ia}} . \tag{4}$$

In practice, the non-negative constraint encourages *additive* combinations of parts; for e.g., in face recognition, additive features include the nose, eyes and mouth. This characteristic often leads to overcomplete bases but also favors components that are relatively localized and easily interpreted over more compact representations. As such, it has been very useful in a variety of applications, examples of which include document clustering [7] and image representation [8].

We believe that this property will in turn be particularly useful in the present context as the factor loadings are equally important for the elucidation of occupational skill dimensions as for the subsequent statistical analysis.

2.4 Computational Considerations

All the analysis in this paper was conducted in the Python and R programming environments. The NMF implementation from the well known *scikit-learn*[2] was used while Factor analysis was performed in R and coordinated from Python using RPY2 (a Python variant provided by scikit-learn was evaluated but was rejected as it lacked a proper varimax implementation).

To retrieve the most representative skills for each factor, a procedure similar to that used in MC (but not identical) was adopted to facilitate comparison. Briefly, this worked as follows:

[2] http://www.scikit-learn.org.

1. All items with loadings below a threshold which was empirically set between the 90th to 95th percentile of factor loadings were retained.
2. All factors with at least three remaining items were retained.
3. Items not loading on any particular factor were discarded.
4. The procedure is iterated until all skills loaded on at least one factor.

While some "manual optimization" proved helpful in improving the clarity of results obtained, our experience was that these results were not overly sensitive to these parameters and were broadly consistent over a range of settings.

3 Results

The methods and approach discussed in the previous sections were applied to the data and the results will now be discussed. Due to space constraints most of the analysis focuses on the 2006 data, while for the regression analysis the 2014 loadings will be used as the dependent variable.

Figure 2 depicts factor loadings extracted from the 2006 data. From these figures, it is apparent that both methods were able to extract loading matrices that concentrated on a small subset of the skills. However, the factors extracted using FA tended to be weighted towards the first few factors, where the loadings were heavily concentrated on a larger number of coefficients, as may be expected based on the properties of the factor analysis algorithm. In contrast, NMF tended to extract factors with more evenly balanced loadings both across and within individual factors, which appears to stem from the lack of an imposed ordering on the factors.

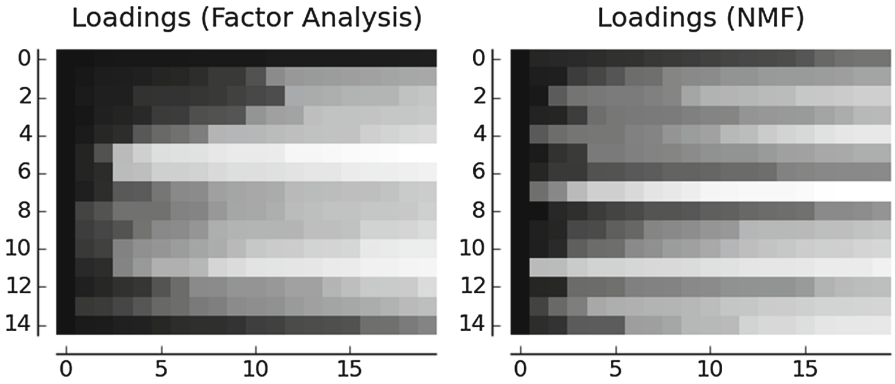

Fig. 2. Loading matrices for first 15 factors extracted using FA and NMF

To better understand the implications of this, a number of factors were extracted from the O*NET occupations data using the procedure described Sect. 2.4. When using FA, the extracted factors were:

1. **Leadership:** Instructing; Learning Strategies; Social Perceptiveness
2. **Manual:** Stamina; Gross Body Coordination; Trunk Strength
3. **Equipment:** Troubleshooting; Repairing; Equipment Maintenance
4. **Vehicle Operation:** Night Vision; Peripheral Vision; Glare Sensitivity
5. **Perception:** Flexibility of Closure; Perceptual Speed; Speed of Closure
6. **Mathematical:** Number Facility; Mathematical Reasoning; Mathematics

Using NMF, the factors extracted were as follows:

1. **Research:** Science; Analyzing Data or Information; Documenting/Recording Information
2. **Computing:** Interacting With Computers; Documenting/Recording Information; Processing Information
3. **Maintenance:** Repairing; Installation; Repairing and Maintaining Mechanical Equipment
4. **Design:** Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment; Mathematics; Technology Design
5. **Perception:** Reaction Time; Hearing Sensitivity; Response Orientation
6. **Leadership:** Training and Teaching Others; Instructing; Coaching and Developing Others
7. **Manual:** Performing General Physical Activities; Stamina; Static Strength

Firstly, it is reassuring to note that the two methods produced factors that are broadly similar (it would be very disconcerting if completely different factors were produced), though there were also some very interesting differences. More comprehensive analysis is required to fully understand these, but it does appear that the NMF factors are more specific and emphasize the occupations (the *Work Activities*), while FA factors are higher level and refer to general categories of abilities (the *Skills* and *Abilities*). As the data consists of vectors of occupations, this is consistent with the main characteristic of NMF which is to produce additive components, where each occupation is composed of multiple activities. The presence of negative coefficients in FA factors would allow more compact basis vectors which, for example, could leverage differences between skills to produce a more compact but higher level basis set which focus on the *requirements* of jobs - i.e. the abilities and skills of ideal candidates. This explains why NMF tended to produce a larger, but generally more specific and more easily interpretable basis set.

Finally, following the methodology in MC, we now study the intensive changes via a linear regression using the 2006 factors as the independent variables, and the 2014 factors as the dependent variables. The resulting regression coefficients are shown in Table 1.

The main observations from Table 1 were as follows:

1. We note that the diagonal elements dominate Table 1, as would be expected. However, there are also substantial off-diagonal elements, which point to skill substitutions within occupations.

Table 1. Changes in Job Characteristics, analyzed using 2006 Factors

	Research	Computing	Maint'nce	Design	Perception	Leadership	Manual
Research	**0.6768**	0.0359	−0.0887	−0.0384	−0.0578	0.2462	−0.0376
Computing	0.0352	**0.7562**	−0.0711	−0.0068	−0.2041	−0.0767	−0.1303
Maintenance	−0.0115	0.0048	**0.6810**	−0.0186	0.1634	−0.0374	0.0167
Design	−0.0492	−0.0116	−0.0714	**0.4865**	−0.0246	0.0455	−0.0387
Machine	0.0074	−0.0001	0.0251	−0.0029	**0.5014**	−0.1054	−0.0782
Leadership	0.0276	−0.0809	−0.0418	0.0076	−0.0787	**0.5900**	−0.0203
Manual	−0.0137	−0.1075	−0.0140	−0.0344	−0.0478	−0.0614	**0.7914**
(Intercept)	−0.0107	0.0750	0.0135	−0.0104	0.1124	0.1029	0.0724
Rsq	0.8131	0.8159	0.7777	0.5605	0.6132	0.6110	0.8244

2. However it is difficult to analyze specific substitutions due to possible correlations between the components. In this aspect NMF is less useful than FA where, by definition, factors start out uncorrelated and any subsequent cross terms can be directly interpreted as the degree of skill substitutions.
3. As was explained in MC, negative intercepts actually imply an *increase* in the importance of particular skills. Here we see that two skills: *Research* and *Design*, both of which are intellectually demanding and resistant to automation, experienced increases in importance.
4. Similarly, the category with the largest decrease in importance was *Perception*. This also corroborates the results in MC, where the Perception component was the most negatively impacted over the same time period.

4 Conclusions and Future Plans

Broadly, the results presented here matched those in MC, which supports the validity and potential usefulness of NMF as a viable alternative to FA for studying the underlying skill dimensions within the O*NET database. However, both methods have their respective strengths and weaknesses. The key findings of this study are that:

1. While not a "drop-in" replacement, NMF provides a valuable addition to the analytical toolkit for studying changes in occupational skill compositions.
2. NMF seems to extract more specific factors which better reflect work activities, while FA tended to extract higher level factors consisting mainly of abilities and skills.
3. However, when performing regression analysis, for e.g. to study intensive changes in skills requirements, NMF factors produced results that were less clear. This could be because of residual correlations within the factors while, on the other hand, FA by definition produces factors which are orthogonal.

Acknowledgement. The authors would like to express their gratitude to the Masdar Institute of Science and Technology for supporting this research.

References

1. Autor, D.H., Dorn, D.: How technology wrecks the middle class. The New York Times, 24 August 2013
2. MacCrory, F., Westerman, G., AlHammadi, Y., Brynjolfsson, E.: Racing with and against the machine: changes in occupational skill composition in an era of rapid technological advance. In: Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems, ICIS 2014, Auckland, New Zealand, 14–17 December 2014
3. Baumol, W.J., Bowen, W.G.: Performing Arts-the Economic Dilemma: A Study of Problems Common to Theatre, Opera, Music and Dance. MIT Press, Cambridge (1966)
4. Acemoglu, D., Autor, D.: Skills, tasks and technologies: implications for employment and earnings. Handb. Labor Econ. **4**, 1043–1171 (2011)
5. Autor, D., Levy, F., Murnane, R.: The skill content of recent technological change: an empirical exploration. Q. J. Econ. **118**(4), 1279–1333 (2003)
6. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, pp. 556–562 (2001)
7. Xu, W., Liu, X., Gong, Y.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, pp. 267–273. ACM (2003)
8. Liu, H., Wu, Z., Li, X., Cai, D., Huang, T.S.: Constrained nonnegative matrix factorization for image representation. IEEE Trans. Pattern Anal. Mach. Intell. **34**(7), 1299–1311 (2012)