Journal of Research in Personality 57 (2015) 100–109

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

A spectral clustering approach to the structure of personality: Contrasting the FFM and HEXACO models

Sean Brocklebank [a,*], Scott Pauls [b], Daniel Rockmore [b,c], Timothy C. Bates [d,*]

[a] School of Economics, University of Edinburgh, United Kingdom
[b] Department of Mathematics, Dartmouth College, Hanover, NH, USA
[c] The Santa Fe Institute, Santa Fe, NM, USA
[d] Department of Psychology, University of Edinburgh, United Kingdom

ARTICLE INFO

Article history:
Available online 22 May 2015

Keywords:
Personality structure
NEO
FFM
5FM
IPIP
HEXACO
Spectral clustering
Factor analysis

ABSTRACT

Alternative analytic methods may help resolve the dimensionality of personality and the content of those dimensions. Here we tested the structure of personality using spectral clustering and conventional factor analysis. Study 1 analysed responses from 20,993 subjects taking the 300-item IPIP NEO personality questionnaire. For factor analysis, a five-factor solution recovered the FFM domains while the six-factor solution yielded only a small and hard to interpret sixth factor. By contrast, spectral clustering analysis yielded six-cluster solutions congruent with the HEXACO model. Study 2 analysed data from 1128 subjects taking the 100-item HEXACO-PI-R. Unambiguous support was found for a six-cluster solution. The psychological content of the 6 clusters and their relationship to the FFM domains is discussed.

1. Introduction

Taxonomy is basic for any science – often being referred as the "facts" of a field, for which theories then compete to account (McCrae & John, 1992). The end of the 20th century saw the emergence of a broad consensus regarding the basic dimensions of personality as consisting of five orthogonal, broad bandwidth domains, with numerous facets clustered beneath these core domains. Variants such as the Five Factor Model (FFM) (McCrae & Costa, 1997, 2003) or the Big Five (Goldberg, 1990) have often been seen as opportunities to refine and redefine the orientation of domains within this space, but both support five basic domains of personality. Considering the question of whether there are additional basic factors of personality not already included among the FFM, McCrae and John (1992, p. 190) concluded that this "*appears increasingly unlikely, given the wealth of data in support of the comprehensiveness of the FFM*". However while the item space of large FFM inventories may be comprehensive, there remains the possibility that the contents of personality may usefully be arranged in terms of other than five basic domains.

One actively researched alternative to the FFM – the HEXACO model – suggests that personality consists of six rather than five basic domains, with the additional domain being one of Honesty–Humility (Ashton & Lee, 2007). Models with fewer domains, for instance Eysenck's PEN structure, retain some support also (e.g. Tiliopoulos, Pallier, & Coxon, 2010). A second area of research focuses on psychological content: the characteristics of high and low poles on each of the basic domains of personality (e.g. Costa & McCrae, 1998). Alternative methods to analyse the structure contained in personality data may yield differential support or elucidate differences between these competing models.

Measuring the structure of personality requires both that the questionnaire or other data adequately sample the universe of personality variation, and that the analytic methods applied to these data are sensitive to this information. The Five Factor Model emerged from a research programme designed, above all, to generate a comprehensive sampling of human personality (McCrae & John, 1992). If the FFM items universe is comprehensive, then adequate analytic tools should then reveal the basic dimensions of personality, even if these are fewer or larger than five. In Studies 1 and 2, we apply one such tool – spectral clustering – to two large (n = 20,993, n = 1126) data sets, one sampling the FFM domains, and one explicitly designed to sample the HEXACO domains.

Before proceeding to these two empirical studies using spectral clustering, we first outline the logic of this method, endeavouring to leverage the reader's existing knowledge of factor analysis.

* Corresponding authors.
E-mail addresses: sean.brocklebank@ed.ac.uk (S. Brocklebank), tim.bates@ed.ac.uk (T.C. Bates).

http://dx.doi.org/10.1016/j.jrp.2015.05.003
0092-6566/

1.1. Spectral clustering

Factor analysis and confirmatory factor analysis (Joreskog, 1969) remain the primary techniques for exploring structure in questionnaire data and are the foundation for the FFM. Of course a set of well known methods have emerged to selecting number of factors in exploratory factor analysis both visually (Cattell, 1966) and analytically (Horn, 1965; Revelle & Rocklin, 1979). A vast range of research has also focussed on contrasting distinct solutions for personality structure via exploratory and confirmatory techniques (e.g. Church & Burke, 1994), and determining criteria for evaluating the structure of personality. Confirmatory factor analytic studies of the FFM have suggested poor fit to the theorised model (Gignac, Bates, & Jang, 2007). Newer exploratory SEM approaches, however, which relax the strict criteria of CFA, allowing an EFA-structure at the item level, indicate that well-fitting five-factor models can be constructed within this framework (Marsh et al., 2010). Interest in exploratory SEM is growing, as is interest in determining criteria for evaluating structure more generally (Hopwood & Donnellan, 2010). Alongside conventional factor analysis-based approaches, alternative analytic strategies widely used outside personality, been used to address the question of the basic structure of personality (Tiliopoulos et al., 2010). In particular, spectral clustering (used in the present report) has emerged as a valuable tool among clustering techniques (Von Luxburg, 2007).

While spectral clustering shares with factor analysis the basic objective of creating a low-dimensional representation of data, it differs substantially in its optimisation target (see e.g., Braun, Leibon, Pauls, & Rockmore, 2011; Leibon, Pauls, Rockmore, & Savell, 2008; Ng, Jordan, & Weiss, 2001). Factor analysis operates on the correlation matrix and tries to replicate it as closely as possible in a smaller number of dimensions (Cattell, 1978; Spearman, 1927). Spectral clustering also has the objective of summarising the data in fewer dimensions, but differs from factor analysis both because it operates on a spatial transformation of the correlation matrix (where each item becomes a point in space) and because it attempts to segregate these items into discrete clusters so that the similar items are kept in the same cluster, while dissimilar items are put into different clusters (see e.g., Braun et al., 2011; Leibon et al., 2008; Ng et al., 2001).

Aspects of spectral clustering have close analogues with factor analysis. For instance, spectral clustering has a cluster-number parameter (k) corresponding to the number of factors requested for extraction in conventional factor analysis. Other aspects, however, are quite distinct. In particular, spectral clustering translates correlations among items into distances (sometimes described as adjacencies – the inverse of distances), and can transform these distances to emphasise particular kinds of relationships by varying what is known as the scale parameter (sigma).

The explanation below is designed to give the casual reader a good intuition about spectral clustering; in addition, online Supplementary Material is provided which contains a technical description, a full list of references and a runnable Matlab implementation of the procedures for those wishing to implement the procedure and/or to learn more about it. Readers interested in the formal mathematical detail should of course consult the source references.

2. Mapping correlations as points in space

The algorithm underlying spectral clustering converts the correlation matrix into a spatial representation – points in space (see Fig. 1). Doing this requires taking the correlation matrix and converting each correlation into a measure of distance (i.e. how far apart is each pair of items?). The measure of distance has two important properties: (i) the measure cannot fall below zero (distances cannot be negative), and (ii) distances are smallest when correlations are largest (so that the items are close together in space), larger for independent items, and largest for negatively linked items.

The conversion of data from a conventional n $\times$ n correlation matrix of variables into spatial distances between these variables is shown in Fig. 1. As shown, items that correlate strongly positively are placed near to each other. Items having a strong negative correlation with each other are placed most far apart. Intermediate correlations translate into intermediate distances. The 3-item case depicted in Fig. 1 is chosen so that each item distance can be realised in the 2-D plane of the page. In the general case, mapping n items can require up to $n - 1$ spatial dimensions. With the translation from correlations into distances achieved, Fig. 2 shows how a spatial representation of items is split into clusters.

The clustering operation (see Fig. 2) creates solutions with k-clusters of items by cutting the paths connecting item-pairs. This generates sets of items each connected to each other within a cluster, but not connected to any item outside the cluster. This is an iterative procedure, which can take a considerable time, as multiple alternative cutting solutions are generated and compared against a criterion, namely to minimise the total length of remaining connections while forming k-clusters of items. This is shown in Fig. 2: It can be seen that this criterion results in the creation of clusters consisting of items that are close to one another, but distant from items in other clusters. We next discuss a feature of spectral clustering which allows it to re-scale the translation of correlations into distances, and which plays an important role in allowing this method to detect structure in data.

3. Scale parameter: sigma

A specific advantage of the spectral clustering algorithm is its inclusion of a scale parameter – sigma (see Fig. 3). By adjusting this parameter, it is possible to vary the relative weight placed on the weakest versus the strongest correlations when performing the optimisation. The scale parameter has some analogues in processes widely used to re-weight correlation matrices in other fields (Sammon, 1969), but has not been deployed in Factor Analysis of personality. Fig. 3 shows this mapping for two example values of sigma. The figure note details how sigma alters the translation of correlation into spatial distance for the example values of sigma.

The ability to reweight correlations is valuable for discovering and understanding structure within data. In particular, setting the scale parameter to a low value emphasises strong correlations among pairs of items. For personality data, the relevance of scale is particularly apparent if a questionnaire is thought to contain a small number of items strongly targeting a domain. For instance, in the present case, we expect a questionnaire designed primarily to assess the FFM domains to contain a relatively small number of Honesty–Humility items. If the HEXACO model is correct these items will nevertheless show strong correlations with each other, and relatively weak correlations with items in the FFM domains. Fig. 4 shows how a low value for sigma can correctly identify clusters represented by only a sparse set of valid items. The low values of sigma increase all inter-item distances but magnify large distances disproportionately more. This has the effect of increasing the distance between valid clusters so they can be more readily identified (this effect of sigma is material for Study 1 below, where we examine personality structure in a questionnaire believed to have sparse coverage of one domain).

When all domains are well represented in the original set of items, changing sigma will not necessarily have any effect on the