Machine Learning Detects Pairwise Associations between SOI and BIS/BAS Subscales, making Correlation Analyses Obsolete

Hermann Prossinger[1,2], Jakub Binter[1,2], Kamila Machová[3], Daniel Říha[2], Silvia Boschetti[2,3]

[1] *Department of Evolutionary Biology, Faculty of Life Sciences, University of Vienna, Vienna, Austria*

[2] *Faculty of Humanities, Charles University, Prague, Czech Republic*

[3] *Faculty of Science, Charles University, Prague, Czech Republic*

ABSTRACT

We use AI techniques to statistically rigorously analyze combinations of query responses of two personality-related questionnaires. One questionnaire probes aspects of a participant's character (SOI) and the other avoidance of aversive outcomes together with approaches to goal orientated outcomes (BIS/BAS). We use one-hot encoding, dimension reduction with a neural network (a seven-layer auto-encoder) and two clustering algorithms to detect associations between the twelve combinations of SOI and BIS/BAS groups. We discover that for most combinations more than one association exists. Traditional, fallacious statistical methods cannot find these outcomes.

Keywords: One-hot Encoding, Autoencoder, Neural Networks, DBSCAN Clustering, Spectral Clustering, BIS/BAS, SOI, Heat Maps

INTRODUCTION

In the fields of psychology, sociology and medicine it is common to use questionnaires to evaluate manifestations of certain traits or states. To do so, researchers habitually rely on aggregations of multiple Likert-type questions to compute an index. Because all responses are categorical/ordinal variables, it is fallacious to convert these to cardinal (computable) integers in a trivial linear mapping. We are thus challenged to analyze the responses in a statistically rigorous manner.

We use two questionnaires:

(a) “Revised Socio-Sexual Orientation (Index)” (SOI-R; Penke, & Asendorpf, 2008). This questionnaire probes aspects of each participant’s tendencies for uncommitted sexual behavior. It has three groups of queries, ‘Attitude’, ‘Behavior’, and ‘Desire’. There are five options for each query (Table 1). (In this paper, we use the shortened label SOI for SOI-R.)

(b) “Behavioral Inhibition and Behavioral Activation Scale” (BIS/BAS; Carver & White, 1994). There are four groups of questions, one probing motivation to avoid aversive outcomes (‘Inhibition’) and three probing the approach to goal-oriented outcomes (‘Fun’ Seeking, ‘Reward’ Responsiveness, and ‘Drive’). Not all groups have the same number of queries. There are four options for each query (Table 1).

Since a linear trivial mapping is not valid, we can neither compute indices nor correlations. Even though these assumptions are violated, correlations are often accepted by reviewers and editors, despite warnings (Hoekstra et al., 2012). We present one modern, rigorous statistical method in the context of our field — sexuality research — that reveals the associations between inhibition, activation and tendency for uncommitted sexual activity.

MATERIALS AND METHODS

891 participants aged 18–50 responded to all queries in the survey. First: for each query, we convert each response to a vector via one-hot encoding (Murphy, 2012). The vectors of the many responses are then concatenated to form a feature vector. Each respondent, therefore, has an associated feature vector of length determined by the total number of options for all the queries in the group combination.

The feature vectors have a (very) high dimension (namely, between 31 and 43, depending on the number of queries in the questionnaire section), so a conventional analysis using SEM or some other model will be invalid because of the 'curse of dimensionality'. Components for each participant's feature vectors are not linearly independent, and the feature vectors of all participants have (nonlinear) interdependencies among one another as well. It is these two interdependencies we are interested in.

Table 1: Listing of the response options for each BIS/BAS category and each SOI category.

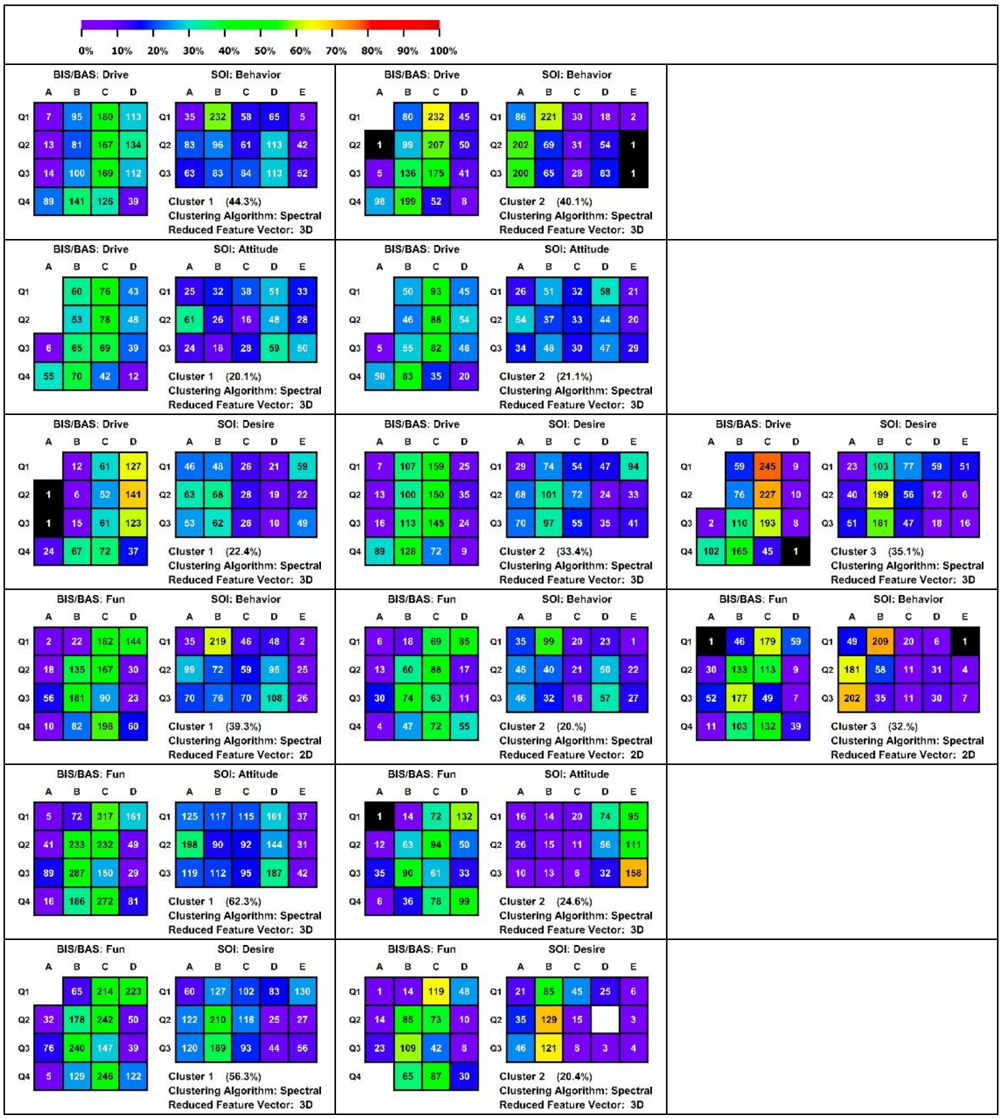
Probe	Ordinal Categories	Response		
		Behavior	Attitude	Desire
SOI	A	Zero	Strongly disagree	Never
	B	One	Disagree	Very seldom
	C	Two to three	Neutral	About once a month
	D	Four to seven	Agree	About once a day
	E	Eight or more	Strongly agree	Nearly every day
BIS/BAS	A	Very false for me		
	B	Somewhat false for me		
	C	Somewhat true for me		
	D	Very true for me		

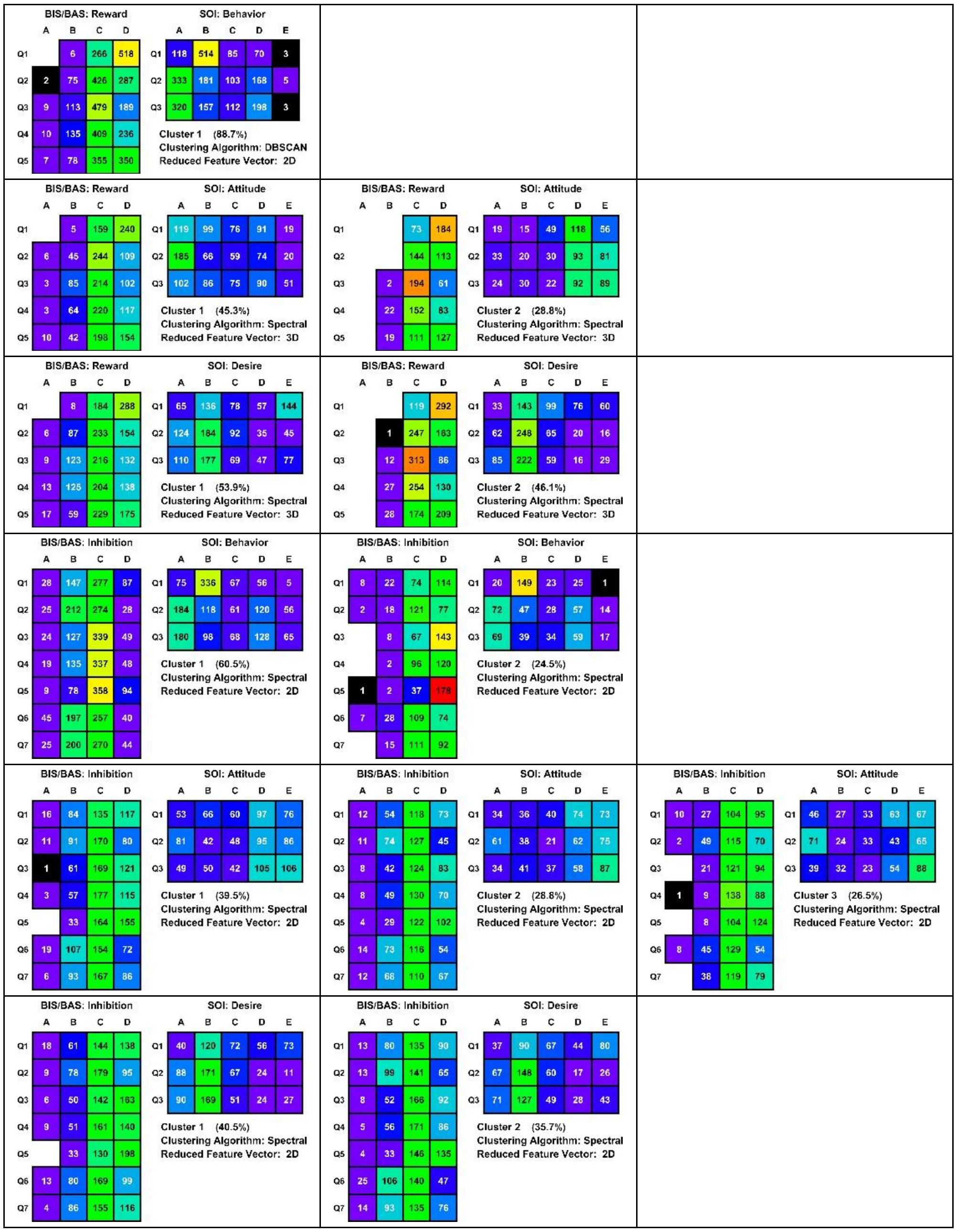
We generate dimension-reduced feature vectors by applying a neural network (specifically: an autoencoder (Goodfellow, 2016) with seven layers), to each combination of SOI and BIS/BAS groups. For every SOI and BIS/BAS combination: (a) We search for clusters in the dimension-reduced feature vector space, using the unsupervised Spectral (Ng et al., 2002) and DBSCAN (Ester et al., 1996) clustering algorithms. We thus have 4 cluster spectra for each combination (each clustering algorithm is applied in 2D and 3D). (b) We determine meaningful associations by choosing membership probabilities above 20% in the clusters of each combination. (c) We "backtrack" for positions of cluster membership to feature vector positions. (d) We construct heat maps of membership percentages in the chosen clusters.

RESULTS

We discover that, for most SOI with BIS/BAS combinations, more than one cluster (and thus more than one association) was detected (Table 2). For each association between a BIS/BAS and SOI combination, cluster membership varies between 41.2% and 100%. Most associations contain 80% to 90% membership (Fig. 1).

Table 2: The heat maps for each association cluster of BIS/BAS responses with SOI responses. Cluster number, probability of the cluster membership, clustering algorithm, and reduced feature vector dimension are listed. Percentage scaling is shown; the colors of the squares in the heat maps are the percentages of the responses for each query separately. The numbers within the squares are the frequencies of members.





We determine the probability of cluster membership by calculating the fractions of member numbers in a cluster relative to the total number of survey participants (891). Because the sample size is large, using the Laplace paradigm for these probability calculations is justified.

Only one combination has only one cluster; it is also the only cluster that was detected using the (unsupervised) DBSCAN clustering algorithm (Ester et al., 1996). All other associations were found with the (unsupervised) Spectral clustering algorithm (Ng et al., 2002). Not all clusters were in three-dimensional reduced feature vector space (Table 2).

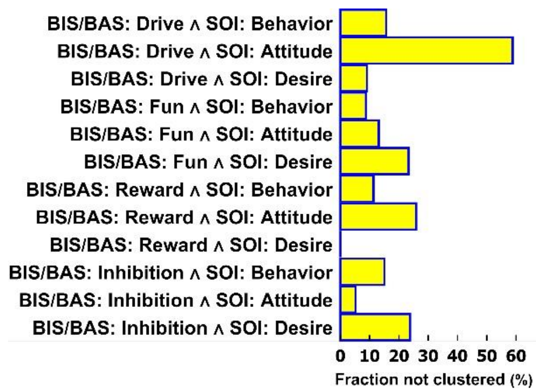


Figure 1. The fractions of membership that are *not* clustered. For one association, all members are in one cluster only (‘Reward’ vs. ‘Behavior’, Table 2). For one association (‘Reward’ vs. ‘Desire’, Table 2), no members are not clustered. For the combination ‘Drive’ vs. ‘Attitude’, membership in any cluster is least.

For some BIS/BAS with SOI combinations, associations occurred with high probability (Table 2). This implies that, for some combinations of BIS/BAS and SOI, associations were very pronounced. Because only one association was limited to only one cluster, we point out that for all the other 11 combinations, more than one association occurred.

Because of space limitations in this paper, we highlight cluster comparisons for only three BIS/BAS and SOI combinations. We specifically avoid psychological interpretations and implications of the highlighting, and focus on comparisons of how responses are skewed in one or both clusters, as discovered by AI.

In the ‘Inhibition’ vs. ‘Behavior’ combination, Cluster 1 has predominantly ‘C’ responses (with a mild contribution of ‘B’) for ‘Inhibition’, while Cluster 2 has predominantly ‘D’ responses for ‘Inhibition’ (especially for query ‘Q5’). In both clusters, the distribution of responses for ‘Behavior’ is close to equal. We conclude that the responses to ‘Inhibition’ queries drive the segmentation of the clusters.

In ‘Fun’ vs. ‘Attitude’, the responses to ‘Fun’ queries in both clusters are close to equal, while the responses to ‘Attitude’ are not. In Cluster 1, responses to SOI queries are skewed towards ‘A’ and ‘B’; in Cluster 2, responses for all three SOI queries are almost exclusively ‘E’. We conclude that the responses to ‘Attitude’ queries drive the segmentation of the clusters.

In ‘Reward’ vs. ‘Attitude’, the responses to ‘Attitude’ queries in both clusters are both skewed to responses ‘C’ and ‘D’, but in Cluster 2 more strongly. The responses to ‘Attitude’ are oppositely skewed: in Cluster 1, towards responses ‘A’, while in Cluster 2 strongly towards ‘D’ and ‘E’. In this BIS/BAS and SOI combination, segmentation between the two clusters is due to a different skewing in both ‘Reward’ and ‘Attitude’.

DISCUSSION

There exist major problems when attempting to make sense of the data obtained from questionnaires that are so often used in sociological, psychological, and medical fields; using correlational analysis is the ‘culprit’. Many experience-oriented sciences are “cursed” (let´s introduce the idiom *curse of psychology*) by the researchers trying to grasp the experiences perceived in others and convert those into queries that (by introducing a further mistake) describe members of a population on a linear scale after creating an index (to aggravate the problem even further, although this mistake had been pointed out 62 years ago; Blalock, 1960). 10 years ago, it was pointed out to psychologists that violating assumptions lead to false conclusions (Hoekstra, 2012). Our approach using AI remedies this.

We avoid the “curse of dimensionality” that occurs when including a large number of queries in questionnaires but not dimension-reducing the response vectors. If not, modeling noise is the result. Higher-dimensional spaces have properties that are counterintuitive and do not allow for easy understanding of the questionnaire outcome. Remarkably, dimension reduction not only prevents modeling noise, but has the fortunate extra feature of offering easier interpretation(s).

Furthermore, by using one-hot encoding, we have avoided the listed mistakes. We reiterate: one-hot encoding (a) does not fallaciously assign cardinal numbers to categorical variables via a linear mapping (Blalock, 1960); (b) ensures that the response vectors for each single question/statement are mutually independent (orthonormal); (c) introduces no a-priori metric; (d) allows for the detection of interdependencies (including nonlinear ones) among the components of the feature vectors; (e) allows for the detection of interdependencies among the feature vectors from the different respondents.

Fallaciously calculating the correlation between any BIS/BAS and SOI combination would yield only one value. It is impossible, therefore, to detect the existence of clusters, which provide evidence that more than one association between a BIS/BAS and SOI combination can exist. The heat maps of the clusters display the characteristics of members of the clusters (such as high frequencies for certain responses to some queries). By taking advantage of AI clustering algorithms, we are able to characterize the fact that more than one association can exist (and does in 11 out of 12 cases; Table 2) within each BIS/BAS and SOI combination.

CONCLUSIONS

By avoiding the fallacies of converting ordinal query responses to computable integers and then using a linear mapping to calculate correlations between combinations of BIS/BAS and SOI combinations, we discover that more than one association can exist between these combinations. (Fallacious calculations cannot discover this, because only one correlation exists for each combination.)

In addition, we discovered that association probabilities are usually high, and, furthermore, that only one combination ('Reward' vs. 'Behavior') has one and only one association, albeit with a high probability.

Detection of more than one association for each combination radically changes the insights regarding personality types that relate character and motivation to each participant's sexuality.

ACKNOWLEDGMENTS

We thank narodnipanel.cz and pokusnikralici.cz for their help with the data collection. Our thanks also go to all participants for taking part in the study. The project is funded by the Czech Science Foundation (Grant No. 19-12885Y "Behavioral and Psycho-Physiological Response on Ambivalent Visual and Auditory Stimuli Presentation"). We thank Lars Penke, Göttingen, for consultation and insightful discussions.

ETHICS STATEMENT

The Helsinki Convention was adhered to at all times, throughout the presenting of the questionnaire, and during the collection of responses. We abided by the wishes of those who did not agree with the terms and conditions (these were not allowed to proceed with the study). All data were anonymized. The study is part of a set of projects and has been approved by Institutional Review Board of the Faculty of Science, Charles University, Prague (#2018/08).

REFERENCES

Blalock Jr., H.M. (1960) *Social Statistics*. McGraw-Hill, New York, USA.

Carver, C.S., White, T.L. (1994) Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: The BIS/BAS scales. *Journal of Personality and Social Psychology*, 67, 319–333.

Ester, M., Kriegel, H-P., Sander, J., Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J., Fayyad, U.M. (eds.). *Proceedings of the Second International*

Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press, 226–231.

Goodfellow, I., Bengio, Y., Courville, A. (2016) *Deep Learning*. The MIT Press, Cambridge, MA, USA.

Hoekstra, R., Kiers, H.A.L., Johnson, A. (2012) Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology* 3 1–9. doi: 10.3389/fpsyg.2012.00137

Murphy K.P. (2012) *Machine Learning. A Probabilistic Perspective*. The MIT Press, Cambridge, MA, USA.

Ng, A.Y., Jordan, M.I., Weiss, Y. (2002) On spectral Clustering: Analysis and an Algorithm. *Advances in Neural Information Processing Systems* 14, 849–856.

Penke, L., Asendorpf, J.B. (2008). Beyond global sociosexual orientations: a more differentiated look at sociosexuality and its effects on courtship and romantic relationships. *Journal of personality and social psychology*, 95, 1113.